PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C07K 14/71

A2

(11) International Publication Number: WO 98/55512

(43) International Publication Date: 10 December 1998 (10.12.98)

(21) International Application Number: PCT/EP98/03193

(22) International Filing Date: 28 May 1998 (28.05.98)

(30) Priority Data:
97201645.5 2 June 1997 (02.06.97) EP
(34) Countries for which the regional or international application was filed: NL et al.

(71) Applicant *(for all designated States except US)*: VLAAMS INTERUNIVERSITAIR INSTITUUT VOOR BIOTECHNOLOGIE [BE/BE]; Rijvisschestraat 118, B–9052 Zwijnaarde (BE).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: VERSCHUEREN, Kristin [BE/BE]; Twee Leeuwenstraat 22, B–3078 Everberg (BE). REMACLE, Jacques [BE/BE]; Avenue des Lilas 7, B–4280 Hannut (BE). HUYLEBROECK, Danny [BE/BE]; Lijsterlaan 15, B–1770 Liede/kerke (BE).

(74) Common Representative: VLAAMS INTERUNIVERSITAIR INSTITUUT VOOR BIOTECHNOLOGIE; Rijvisschestraat 118, B–9052 Zwijnaarde (BE).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report

(54) Title: SMAD–INTERACTING POLYPEPTIDES AND THEIR USE

(57) Abstract

The current invention concerns SMAD–interacting protein(s) obtainable by a two–hybrid screening assay whereby Smad1 C–domain fused to GAL4 DNA–binding domain as bait and a cDNA library from mouse embryo as prey are used. Some characteristics of a specific SMAD interacting protein so–called SIP1 are the follwing: a) it fails to interact with full size XSmad1 in yeast; b) it is a member of the family of zinc finger/homeodomain proteins including δ–crystallin enhancer binding protein and/or Drosophila zfh–1; c) $SIP1_{czf}$ binds to E2 box sites, d) $SIP1_{czf}$ binds to the Brachyury protein binding site; e) it interferes with Brachyury–mediated transcription activation in cells and f) it interacts with C–domain of Smad 1,2 and 5. The minimal length of the amino acid sequence necessary for binding with Smad appears to be a 51 aa domain encompassing aa 166–216 of SEQ ID NO 2 having the amino acid sequence as depicted in the one letter code: QHLGVGMEAPLLGFPTMNSNLSEVQKVLQIVDNTVSRQKMDCKTEDISKLK.

Smad-interacting polypeptides and their use

The present invention relates to Smad - interacting polypeptides (so-called SIP's) such as cofactors for Smad proteins and the use thereof.
The development from a single cell to a fully organized organism is a complex process wherein cell division and differentiation are involved. Certain proteins play a central role in this process. These proteins are divided into different families of which the transforming growth factor β (TGF-β) family of ligands, their serine/threonine kinase (STK) receptors and their signalling components are undoubtedly key regulatory polypeptides. Members of the TGF-β superfamily have been documented to play crucial roles in early developmental events such as mesoderm formation and gastrulation, but also at later stages in processes such as neurogenesis, organogenesis, apoptosis and establishment of left-right asymmetry. In addition, TGF-β ligands and components of their signal transduction pathway have been identified as putative tumor suppressors in the adult organism.
Recently, Smad proteins have been identified as downstream targets of the serine/threonine kinase (STK) receptors (Massagué,1996, Cell,85, p. 947-950). These Smad proteins are signal transducers which become phosphorylated by activated type I receptors and thereupon accumulate in the nucleus where they may be involved in transcriptional activation. Smad proteins comprise a family of at least 5 subgroups which show high cross-species homology. They are proteins of about 450 amino acids (50-60kDa) with highly conserved N-terminal and C-terminal domains linked by a variable, proline-rich, middle region. On the basis of experiments carried out in cell lines or in Xenopus embryos, it has been suggested that the subgroups define distinct signalling pathways: Smad1 mediates BMP2/4 pathways, while Smad2 and Smad3 act in TGF-β / activin signal transduction cascades. It has been demonstrated that these Smads act in a complex with Smad4 (dpc-4) to elicit certain activin, bone morphogenetic protein (BMP) or TGF-β

responses (Lagna et al., 1996, Nature, 383, p.832-836 and Zhang et al., 1996, Nature,383, p.168-172).

Smad proteins have a three-domain structure and their highly conserved carboxyl domain (C-domain) is necessary and sufficient for Smad function in the nucleus. The concept that this domain of Smad proteins might interact with transcription factors in order to regulate transcription of target genes has previously been put forward (Meersseman et al, 1997, Mech.Dev., 61, p.127-140). This hypothesis has been supported by the recent identification of a new winged-helix transcription factor (FAST1) which forms an activin-dependent complex with Smad2 and binds to an activin responsive element in the Mix-2 promotor (Chen et al. , Nature 383, p. 691-696, 1996). However, cofactors for Smad proteins other than FAST 1 have not been identified yet.

Beyond the determination of the mechanism of activation of Ser/Thr kinase receptors and Smad, and the heteromerization of the latter, little is known about other downstream components in the signal transduction machinery. Thus, understanding how cells respond to TGF-β related ligands remains a crucial central question in this field.

In order to clearly demonstrate that Smad proteins might have a function in transcriptional regulation -either directly or indirectly- it is necessary to identify putative co-factors of Smad proteins, response elements in target genes for these Smad proteins and/or co-factors, and to investigate the ligand-dependency of these activities.

To understand those interactions molecular and developmental biology research on (i) functional aspects of the ligands, receptors and signaling components (in particular members of the Smad family), in embryogenesis and disease, (ii) structure-function analysis of the ligands and the receptors, (iii) the elucidation of signal transduction, (iv) the identification of cofactors for Smad (related) proteins and (v) ligand-responsive genes in cultured cell and the Drosophila, amphibian, fish and murine embryo are all of utmost importance.

It is our invention that by carrying out a two hybrid screening assay (Chien et al., 1991, PNAS,88, p.9578-9582) SMAD interacting protein(s) are obtainable whereby Smad C-domain fused to a DNA-binding domain as bait and a vertebrate cDNA library as prey respectively are used. It is evident for those skilled in the art that other appropriate cDNA libraries can be used as well. By using for instance Smad1 C-domain fused to GAL4 DNA-binding domain and a mouse embryo cDNA as bait and prey respectively, a partial Smad4 and other Smad-interacting protein (SIP) cDNAs, including SIP1, were obtained.

Surprisingly it has been found that at least four SMAD interacting proteins thus obtained contain a DNA binding zinc finger domain. One of these proteins, SIP1, is a novel member of the family of zinc finger/homeodomain proteins containing δ-crystallin enhancer binding protein and certain *Drosophila* zfh-1, the former of which has been identified as a DNA-binding repressor. It has been shown that one DNA binding domain of SIP1 (the C-terminal zinc finger cluster or SIP1$_{czf}$) binds to E2 box regulatory sequences and to the *Brachyury* protein binding site. It has been demonstrated in cells that SIP1 interferes with E2 box and *Brachyury*-mediated transcription activation. SIP1 fails to interact with full-size Smad in yeast. It is shown for the first time that Smad proteins can interact with a DNA-binding repressor and as such may be directly involved in TGF-ß ligand-controlled repression of target genes which are involved in the strict regulation of normal early development.

In summary some characteristics of SIP 1 are the following:

a) it fails to interact with full size XSmad1 in yeast
b) it is a new member of the family of zinc finger/homeodomain proteins including δ-crystallin enhancer binding protein and/or Drosophila zfh-1
c) SIP1$_{czf}$ binds to E2 box sites
d) SIP1$_{czf}$ binds to the Brachyury protein binding site
e) it interferes with Brachyury-mediated transcription activation in cells and
f) it interacts with C-domain of Smad 1, 2 and/or 5

With E2 box sites is meant a -CACCTG- regulatory conserved nucleotide sequence which contains the binding site CACCT for δ-crystallin enhancer binding proteins as described in Sekido et al, 1996, Gene, 173, p.227-232.
These E2 box sites are known targets for important basic helix-loop-helix (bHLH) factors such as MyoD , a transcription factor in embryogenesis and myogenesis.

So, the SIP1 according to the invention (a zinc finger/homeodomain protein) binds to specific sites in the promoter region of a number of genes which are relevant for the immune response and early embryogenesis and as such may be involved in transcriptional regulation of important differentiation genes in significant biological processes such as cell growth and differentiation, embryogenesis, and abnormal cell growth including cancer.

Part of the invention is also an isolated nucleic acid sequence comprising the nucleotide sequence as provided in SEQ ID NO 1 coding for a SMAD interacting protein or a functional fragment thereof.
Furthermore a recombinant expression vector comprising said isolated nucleic acid sequence (in sense or anti-sense orientation) operably linked to a suitable control sequence belongs to the present invention and cells transfected or transduced with a recombinant expression vector as well.
The current invention is not limited to the exact isolated nucleic acid sequence comprising the nucleotide sequence as mentioned in SEQ ID NO 1 but also a nucleic acid sequence hybridizing to said nucleotide sequence as provided in SEQ ID NO 1 or a functional part thereof and encoding a Smad interacting protein or a functional fragment thereof belongs to the present invention.

To clarify with "hybridization" is meant conventional hybridization conditions known to the skilled person, preferably appropriate stringent hybridization conditions. Hybridization techniques for determining the complementarity of nucleic acid sequences are known in the art.
The stringency of hybridization is determined by a number of factors during hybridization including temperature, ionic strength, length of time and composition

of the hybridization buffer. These factors are outlined in, for example, Maniatis et al. (1982) Molecular Cloning; A laboratory manual (Cold Spring Harbor Press, Cold Spring Harbor, N.Y.).

Another aspect of the invention is a polypeptide comprising the amino acid sequence according to SEQ.ID.NO 2 or a functional fragment thereof.

To the scope of the present invention also belong variants or homologues of amino acids enclosed in the polypeptide wherein said amino acids are modified and/or substituted by other amino acids obvious for a person skilled in the art. For example post-expression modifications of the polypeptide such as phosphorylations are not excluded from the scope of the current invention.

The polypeptide or fragments thereof are not necessarily translated from the nucleic acid sequence according to the invention but may be generated in any manner, including for example, chemical synthesis or expression in a recombinant expression system. Generally "polypeptide" refers to a polymer of amino acids and does not refer to a specific length of the molecule. Thus, linear peptides, cyclic or branched peptides, peptides with non-natural or non-standard amino acids such as D-amino acids, ornithine and the like, oligopeptides and proteins are all included within the definition of polypeptide.

The terms "protein" and "polypeptide" used in this application are interchangeable. "Polypeptide" as mentioned above refers to a polymer of amino acids (amino acid sequence) and does not refer to a specific length of the molecule. Thus peptides and oligopeptides are included within the definition of polypeptide. This term does also refer to or include post-translational modifications of the polypeptide, for example, glycosylations, acetylations, phosphorylations and the like. Included within the definition are, for example, polypeptides containing one or more analogs of an amino acid (including, for example, unnatural amino acids, etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

"Control sequence" refers to regulatory DNA sequences which are necessary to affect the expression of coding sequences to which they are ligated. The nature of

such control sequences differs depending upon the host organism. In prokaryotes, control sequences generally include promoter, ribosomal binding site, and terminators. In eukaryotes generally control sequences include promoters, terminators and, in some instances, enhancers, transactivators, transcription factors or 5' and 3' untranslated cDNA sequences. The term "control sequence" is intended to include, at a minimum, all components the presence of which are necessary for expression, and may also include additional advantageous components.

"Operably linked" refers to a juxtaposition wherein the components so described are in a relationship permitting them to function in their intended manner. A control sequence "operably linked" to a coding sequence is ligated in such a way that expression of the coding sequence is achieved under conditions compatible with the control sequences. In case the control sequence is a promoter, it is obvious for a skilled person that double-stranded nucleic acid is used.

“Fragment of a sequence” or “part of a sequence” means a truncated sequence of the original sequence referred to. The truncated sequence (nucleic acid or protein sequence) can vary widely in length; the minimum size being a sequence of sufficient size to provide a sequence with at least a comparable function and/or activity of the original sequence referred to, while the maximum size is not critical. In some applications, the maximum size usually is not substantially greater than that required to provide the desired activity and/or function(s) of the original sequence. Typically, the truncated amino acid sequence will range from about 5 to about 60 amino acids in length. More typically, however, the sequence will be a maximum of about 50 amino acids in length, preferably a maximum of about 30 amino acids. It is usually desirable to select sequences of at least about 10, 12 or 15 amino acids, up to a maximum of about 20 or 25 amino acids.

A pharmaceutical composition comprising above mentioned nucleic acid(s) or a pharmaceutical composition comprising said polypeptide(s) are another aspect of the invention. The nucleic acid and/or polypeptide according to the invention can be optionally used for appropriate gene therapy purposes.

In addition, a method for diagnosing, prognosis and/or follow-up of a disease or disorder by using the nucleic acid(s) according to the invention or by using the polypeptide(s) also form an important aspect of the current invention.

Furthermore in the method for diagnosing, prognosis and/or follow-up of a disease or disorder an antibody ,directed against a polypeptide or fragment thereof according to the current invention, can also be conveniently used. As used herein, the term "antibody" refers, without limitation, to preferably purified polyclonal antibodies or monoclonal antibodies, altered antibodies, univalent antibodies, Fab proteins, single domain antibodies or chimeric antibodies. In many cases, the binding phenomena of antibodies to antigens is equivalent to other ligand/anti-ligand binding.

The term "antigen" refers to a polypeptide or group of peptides which comprise at least one epitope. "Epitope" refers to an antibody binding site usually defined by a polypeptide comprising 3 amino acids in a spatial conformation which is unique to the epitope, generally an epitope consists of at least 5 such amino acids and more usually of at least 8-10 such amino acids.

A diagnostic kit comprising a nucleic acid(s) sequence and/or a polypeptide(s) or antibodies directed against the polypeptide or fragment thereof according to the invention for performing above mentioned method for diagnosing a disease or disorder clearly belong to the invention as well.

Diseases or disorders in this respect are for instance related to cancer, malformation, immune or neural diseases, or bone metabolism related diseases or disorders. In addition a disease affecting organs like skin, lung, kidney, pancreas, stomach, gonad, muscle or intestine can be diagnosed as well using the diagnostic kit according to the invention.

Using the nucleic acid sequences of the invention as a basis, oligomers of approximately 8 nucleotides or more can be prepared, either by excision or synthetically, which hybridize for instance with a sequence coding for SIP or a functional part thereof and are thus useful in identification of SIP in diseased individuals. The so-called probes are of a length which allows the detection of unique sequences of the compound to detect or determine by hybridization as

defined above. While 6-8 nucleotides may be a workable length, sequences of about 10 -12 nucleotides are preferred, and about 20 nucleotides appears optimal. The nucleotide sequence may be labeled for example with a radioactive compound, biotin, enzyme, dye stuff or metal sol , fluorescent or chemiluminescent compound. The probes can be packaged into diagnostic kits. Diagnostic kits include the probe nucleotide sequence, which may be labeled; alternatively, said probe may be unlabeled and the ingredients for labeling may be included in the kit in separate containers so that said probe can optionally be labeled. The kit may also contain other suitably packaged reagents and materials needed for the particular hybridization protocol, for example, standards, wash buffers, as well as instructions for conducting the test.

The diagnostic kit may comprise an antibody, as defined above, directed to a polypeptide or fragment thereof according to the invention in order to set up an immunoassay. Design of the immunoassay is subject to a great deal of variation, and the variety of these are known in the art. Immunoassays may be based, for example, upon competition, or direct reaction, or sandwich type assays.

An important aspect of the present invention is the development of a method of screening for compounds (chemically synthesized or available from natural sources) which affect the interaction between SMAD and SIP's having the current knowledge of the SMAD interacting polypeptides (so called SIP's such as SIP1 or SIP2 as specifically disclosed herein).

A transgenic animal harbouring the nucleic acid(s) according to the invention in its genome also belong to the scope of this invention.

Said transgenic animal can be used for testing medicaments and therapy models as well.

With transgenic animal is meant a non-human animal which have incorporated a foreign gene (called transgene) into their genome; because this gene is present in germ line tissues, it is passed from parent to offspring establishing lines of transgenic animals from a first founder animal. As such transgenic animals are recognized as specific species variants or strains, following the introduction and

integration of new gene(s) into their genome. The term "transgenic" has been extended to chimeric or "knockout" animals in which gene(s), or part of genes, have been selectively disrupted or removed from the host genome.

Depending on the purpose of the gene transfer study, transgenes can be grouped into three main types: *gain-of-function*, *reporter function* and *loss-of-function*.

The *gain-of-function* transgenes are designed to add new functions to the transgenic individuals or to facilitate the identification of the transgenic individuals if the genes are expressed properly (including in some cell types only) in the transgenic individuals.

The *reporter gene* is commonly used to identify the success of a gene transfer effort. Bacterial chloramphenicol acetyltransferase (CAT), β-galactosidase or luciferase genes fused to functional promoters represent one type of *reporter function* transgene.

The *loss-of-function* transgenes are constructed for interfering with the expression of host genes. These genes might encode an antisense RNA to interfere with the posttranscriptional process or translation of endogenous mRNAs. Alternatively, these genes might encode a catalytic RNA (a ribozyme) that can cleave specific mRNAs and thereby cancel the production of the normal gene product.

Optionally loss of function transgenes can also be obtained by over-expression of dominant-negative variants that interfere with activity of the endogenous protein or by targeted inactivation of a gene , or parts of a gene, in which usually (at least a part of) the DNA is deleted and replaced with foreign DNA by homologous recombination. This foreign DNA usually contains an expression cassette for a selectable marker and/or reporter.

It will be appreciated that when a nucleic acid construct is introduced into an animal to make it transgenic the nucleic acid may not necessarily remain in the form as introduced.

By "offspring" is meant any product of the mating of the transgenic animal whether or not with another transgenic animal, provided that the offspring carries the transgene.

To the scope of the current invention also belongs a SMAD interacting protein characterized in that:

a) it interacts with full size XSmad1 in yeast

b) it is a member of a family of proteins which contain a cluster of 5 CCCH-type zinc fingers including Drosophila "Clipper" and Zebrafish "No arches"

c) it binds single or double stranded DNA

d) it has an RNase activity

e) it interacts with C-domain of Smad1, 2 and/or 5.

Part of the invention is also a method for post-transcriptional regulation of gene expression by members of the TGF-β superfamily by manipulation or modulation of the interaction between Smad function and/or activity and mRNA stability.

The current invention is further described in detail hereunder for sake of clarity.

<u>Yeast two-hybrid cloning of Smad-interacting proteins</u>

In order to identify cofactors for Smad1, a two-hybrid screening in yeast was carried out using the *X*Smad1 C-domain fused to GAL4 DNA-binding domain ($GAL4_{DBD}$) as bait, and a cDNA library from mouse embryo (12.5 dpc) as a source of candidate preys. The $GAL4_{DBD}$-Smad1 bait protein failed to induce in the reporter yeast strain GAL4-dependent *HIS3* and *LacZ* transcription on its own or in conjunction with an empty prey plasmid. Screening of 4 million yeast transformants identified about 500 colonies expressing *HIS3* and *LacZ*. The colonies displaying a phenotype which was dependent on expression of both the prey and the bait cDNAs, were then characterized. Plasmids were rescued and the prey cDNAs sequenced (SEQ ID NO's 1-20 of the Sequence Listing enclosed; for each nucleic acid sequence only one strand is depicted in the Listing). Four of these (th1, th12, th76 and th74 respectively also denominated in this application as SIP1, SIP2, SIP5 and SIP7 respectively) are disclosed in detail (embedded in SEQ ID NO 1, 2, 3, 4, 10 and 8 respectively). One (th72= combined SEQ ID NO 6 and 7) encodes a protein in which the GAL4 transactivation domain ($GAL4_{TAD}$) is fused in-frame to a partial Smad4 cDNA, which starts at amino acid (aa) 252 in the proline-rich domain. Smad4 has been shown to interact with other Smad proteins, but no Smad has been picked-up thusfar in a two-hybrid screen in yeast, using the C-domain of another Smad as bait.

These data suggest that the N-domain of both interacting Smad proteins, as well as part of (Smad4) or the entire (Smad1) proline-rich domain, is dispensable for heterodimeric interaction between Smad proteins, at least when using a two-hybrid assay in yeast.

The cDNA insert of the second positive prey plasmid, th1 (embedded in SEQ ID NO 1), encodes a protein in which the $GAL4_{TAD}$-coding sequence is fused in-frame to about a 1.9 kb-long th1 cDNA, which encodes a polypeptide SIP1 (Th1) of 626 aa. Data base searches revealed that SIP1 (Th1) contained a homeodomain-like segment, and represents a novel member of a family of DNA-binding proteins including vertebrate δ-crystallin enhancer binding proteins (δ-EF1) and *Drosophila* zfh-1. These zinc finger/ homeodomain-containing transcription factors are involved in organogenesis in mesodermal tissues and/or development of the nervous system. The protein encoded by th1 cDNA is a Smad interacting protein (SIP) and was named SIP1 (TH1).

SIP1

Characterization of SIP1-Smad interaction in yeast and *in vitro*

The binding of SIP1 (TH1) to full-size *X*Smad1 and modified C-domains was tested. The latter have either an amino acid substitution (G418S) or a deletion of the last 43 aa (Δ424-466). The first renders the Smad homolog in *Drosophila* Mad inactive and abolishes BMP-dependent phosphorylation of Smad1 in mammalian cells. A truncated Mad, similar to mutant Δ424-466, causes mutant phenotypes in *Drosophila*, while a similar truncation in Smad4 (dpc-4) in a loss-of-heterozygosity background is associated with pancreatic carcinomas. SIP1 (TH1) does neither interact with full-size *X*Smad1, nor with mutant Δ424-466. The absence of any detectable association of full-size *X*Smad1 was not due to inefficient expression of the latter in yeast, since one other Smad-interacting prey (th12) efficiently interacted with the full-length Smad bait. Lack of association of SIP1 (TH1) with full-size *X*Smad1 in yeast follows previous suggestions that the activity of the Smad C-domain is repressed by the N-domain, and that this repression is eliminated in mammalian cells by incoming BMP signals. The G418S mutation in the C-domain of Smad 1 does not abolish interaction with SIP1, suggesting that this mutation affects

another aspect of Smad1 function. The ability of the full-size G418S Smad protein to become functional by activated receptor STK activity may thus be affected, but not the ability of the G418S C-domain to interact with downstream targets. This indicates that activation of Smad is a prerequisite for and precedes interaction with targets such as SIP1. The deletion in mutant Δ424-466 includes three conserved and functionally important serines at the C-terminus of Smad which are direct targets for phosphorylation by the activated type I STK receptor.

The C-domains of Smad1 and Smad2 induce ventral or dorsal mesoderm, respectively, when overexpressed individually in *Xenopus* embryos, despite their very high degree of sequence conservation. Very recently, Smad5 has been shown to induce ventral fates in the Xenopus embryo. To investigate whether the striking differences in biological activity of Smad1, -5 and Smad2 could be due to distinct interactions with cofactors, the ability of SIP1 (TH1) protein to interact with the C-domains of Smad1, -5 and Smad2 in a yeast two-hybrid assay was tested. SIP1 (TH1) was found to interact in yeast with the C-domain of all three Smad members. Then the interaction of SIP1 with different Smad C-domains *in vitro* was investigated, using glutathione-S-transferase (GST) pull-down assays. GST-Smad fusion proteins were produced in *E. Coli* and coupled to glutathione-Sepharose beads. An unrelated GST fusion protein and unfused GST were used as negative controls. Radio-labeled, epitope-tagged SIP1 protein was successfully produced in mammalian cells using a vaccinina virus (T7VV)-based system. Using GST-Smad beads, this SIP1 protein was pulled down from cell lysates, and its identity was confirmed by Western blotting. Again, as in yeast, it was found that SIP1 is a common binding protein for different Smad C-domains, suggesting that SIP1 might mediate common responses of cells to different members of the TGF-ß superfamily. Alternatively, Smad proteins may have different affinities for SIP1 *in vivo*, or other mechanisms might determine the specificity, if any, of Smad-SIP1 interaction.

__SIP1 is a new member of zinc finger/homeodomain proteins of the δEF-1 family__

Additional SIP1 open reading frame sequences were obtained by a combination of cDNA library screening with 5'RACE-PCR. The screening yielded a 3.2 kb-long SIP1 cDNA (tw6), which overlaps partially with th1 cDNA. The open reading frame of SIP1

protein encodes 944 aa (SEQ ID NO 2), and showed homology to certain regions in δ-EF1, ZEB, AREB6, BZP and zfh-1 proteins, and strikingly similar organisation of putative functional domains. Like these proteins, SIP1 contains two zinc finger clusters separated by a homeodomain and a glutamic acid-rich domain. Detailed comparisons reveal that SIP1 is a novel and divergent member of the two-handed zinc finger/homeodomain proteins. As in δ-EF1, three of the five residues that are conserved in helix 3 and 4 of all canonical homeodomains are not present in SIP1. SIP1 (Th1) which contains the homeodomain but lacks the C-terminal zinc finger cluster and glutamic acid-rich sequence, interacts with Smad. This interaction is maintained upon removal of the homeodomain-like domain, indicating that a segment encoding aa 44-236 of SIP1 (numbering according to SEQ.ID.NO.2) is sufficient for interaction with Smad. To narrow this domain further down, progressive deletion mutants, starting from the N-terminus, as well as the C-terminus of this 193 aa region were made. Progressive 20 aa deletion constructs were generated by PCR. Two restriction sites (5' end SmaI site, 3' end XhoI site) were built in to allow cloning of amplified sequences in the yeast two hybrid bait vector pACT2 (Clontech). An extensive two hybrid experiment was performed with these so-called SBD mutant constructs as a prey and the *X*Smad1 C-domain as bait. The mutant SBD constructs that encoded aa 166-236 (of SEQ ID NO 2) or aa 44-216 were still able to interact with the bait plasmid, whereas mutant constructs encoding aa 186-236 or aa 44-196 could not interact with the bait. In this way, the smallest domain that still interacts with the *X*Smad1 C-domain was defined as a 51 aa domain encompassing aa 166-216 of SEQ ID NO 2.

The amino acid sequence of said SBD, necessary for the interaction with Smad, thus is (depicted in the one-letter code):

QHLGVGMEAPLLGFPTMNSNLSEVQKVLQIVDNTVSRQKMDCKTEDISKLK

Deletion of an additional 20 aa at the N-or C-terminal end of this region disrupted the Smad binding activity. Subsequently, this 51aa region was deleted in the context of SIP1 protein, again using a PCR based approach, generating an NcoI restriction site at the position of the deletion. This SIP1ΔSBD51 was not able to interact with the

Smad C-domain any longer, as assayed by a "mammalian pull down assay". In these experiments, SIP1, myc-tagged at its N-terminal end was expressed in COS-1 cells together with a GST-*X*Smad1 C-domain fusion protein. Myc-SIP1 protein was co-purified from cell extracts with the GST-*X*Smad1 C-domain fusion protein using gluthatione-sepharose beads, as was demonstrated by Western blotting using anti-myc antibody. Deletion of the 51 aa in SIP1 abolished the interaction, as detected in this assay, with the *X*Smad1 C-domain. (see figure 1).

Analysis of the DNA-binding activity of the C-terminal zinc finger cluster of SIP1.

δ-EF1 is a repressor that regulate the enhancer activity of certain genes. This repressor binds to the E2 box sequence (5'-CACCTG) which is also a binding site for a subgroup of basic helix-loop-helix (bHLH) activators (Sekido, R et al., 1994, Mol.Cell.Biol.,14, p.5692-5700). Interestingly, the CACCT sequence which has been shown to bind δ-EF1 is also part of the consensus binding site for Bra protein. It has been proposed that cell type-specific gene expression is accomplished by competitive binding to CACCT sequences between repressors and activators. δ-EF1 mediated repression could be the primary mechanism for silencing the IgH enhancer in non-B cells. δ-EF1 is also present in B-cells, but is counteracted by E2A, a bHLH factor specific for B-cells. Similarly, δ-EF1 represses the Igκ enhancer where it competes for binding with bHLH factor E47.

The C-terminal zinc finger cluster of δEF-1 is responsible for binding to E2 box sequences and for competition with activators. Considering the high sequence similarities in this region between SIP1 and δ-EF1, it was decided to test first whether both proteins have similar DNA binding specificities, using gel retardation assays. Therefore, the DNA-binding properties of the C-terminal zinc finger cluster of SIP1 (named $SIP1_{CZF}$) was analyzed. $SIP1_{CZF}$ was efficiently produced in and purified from *E. coli* as a short GST fusion protein. Larger GST-SIP1 fusion proteins were subject to proteolytic degradation in *E. coli* .

Purified GST-$SIP1_{CZF}$ was shown to bind to the E2 box of the IgH κE2 enhancer. A mutation of this site (Mut1), which was shown previously to affect the binding of the bHLH factor E47 but not δ-EF1, did not affect binding of $SIP1_{CZF}$. Two

other mutations in this κE2 site (Mut2 and Mut4, respectively) have been shown to abolish binding of δ-EF1 (Sekido *et al.*, 1994) and did so in the case of $SIP1_{CZF}$. In addition, also the binding of $SIP1_{CZF}$ to the Nil-2A binding site of the interleukin-2 promoter, the Bra protein binding site and the AREB6 binding site were demonstrated. The specificity of the binding of $SIP1_{CZF}$ to the Bra binding site was further demonstrated in competition experiments. Binding of $SIP1_{CZF}$ to this site was competed by excess unlabeled Bra binding site probe, while κE2 wild type probe competes, albeit less efficiently than its variant Mut1, which is a very strong competitor. κE2-Mut2 and κE2-Mut4 failed to compete, as did the GATA-2 probe, while the AREB6 site competed very efficiently. From these experiments can be concluded that GST-$SIP1_{CZF}$ fusion protein displays the same DNA binding specificity as other GST fusion proteins made with the CZF region of δ-EF1 and related proteins (Sekido *et al.*, 1994). In addition, it was demonstrated for the first time that SIP1 binds specifically to regulatory sequences that are also target sites for Bra. This may be the case for the other δ-EF1-related proteins as well and these may interfere with Bra-dependent gene activation *in vivo*.

Analyses were done to sites recognized by the bHLH factor MyoD. MyoD has been shown to activate transcription from the muscle creatine kinase (MCK) promoter by binding to E2 box sequences (Weintraub et al., 1994, Genes Dev.,8, p.2203-2211; Katagiri *et al.*, 1997, Exp.Cell Res. 230, p. 342-351). Interestingly, δ-EF1 has also been demonstrated to repress MyoD-dependent activation of the muscle creatine kinase enhancer, as well as myogenesis in 10T½ cells, and this is thought to involve E2 boxes (Sekido *et al.*, 1994). In addition, TGF-ß and BMP-2 have been reported to downregulate the activity of muscle-specific promoters, and this inhibitory effect is mediated by E2 boxes (Katagiri *et al.*, 1997). The latter are present in the regulatory regions of many muscle-specific genes, are required for muscle-specific expression, and are optimally recognized by heterodimers between myogenic bHLH proteins (of the MyoD family) and of widely expressed factors like E47. $SIP1_{CZF}$ was able to bind to a probe that encompasses the MCK enhancer E2 box and this complex was competed by the E2 box oligonucleotide and by other SIP1 binding sites. In addition, a point mutation within this E2 box that is similar to the previously used κE2-Mut4 site also abolished binding of $SIP1_{czf}$. These results

confirm that $SIP1_{czf}$ binds to the E2 box of the MCK promoter. SIP1, as Smad-interacting and MCK E2 box binding protein, may therefore represent the factor that mediates the TGF-ß and BMP repression of the MyoD-regulated MCK promoter (Katagiri *et al.*, 1997).

SIP1 is a BMP-dependent repressor of Bra activator

The experiments have demonstrated that $SIP1_{CZF}$ binds to the Bra protein binding site, IL-2 promoter, and to E2 boxes, the latter being implicated in BMP or TGF-ß-mediated repression of muscle-specific genes. These observations prompted therefore to test whether SIP1 (as $SIP1_{TW6}$) is a BMP-regulated repressor. A reporter plasmid containing a SIP1 binding site (the Bra protein binding site) fused to the luciferase gene was constructed. COS cells, maintained in low serum (0.2%) medium during the transfection, were used in subsequent transient transfection experiments since they have been documented to express BMP receptors and support signaling (Hoodless *et al.*, 1996,Cell, 85, p.489-500). It was found in the experiment that $SIP1_{TW6}$ is not able to change the transactivation activity of Bra protein via the Bra binding site. In addition, no transactivation of this reporter plasmid by $SIP1_{TW6}$ could be detected in the presence of 10% or 0.2% serum, and in the absence of Bra expression vector.

Therefore, identical experiments were carried out in which the cells were exposed to BMP-4. $SIP1_{TW6}$ repressed the Bra-mediated activation of the reporter. It does this in a dose-dependent fashion (amount of $SIP1_{TW6}$ plasmid, concentration of BMP-4). Total repression has not been obtained in this type of experiment, because the transfected COS cells were exposed only after 24 hours to BMP-4. Consequently, luciferase mRNA and protein accumulate during the first 24 hours of the experiment as the result of Brachyury activity. The conclusion from these experiments clearly shows that SIP1 is a repressor of Bra activator, and its activity as repressor is detected only in the presence of BMP. It is important that SIP1 has not been found to be an activator of transcription via Bra target sites. This is interesting, since the presence in δ-EF1-like proteins of a polyglutamic acid-rich stretch (which is also present in $SIP1_{TW6}$ used here) has led previously to the speculation that these repressors might act as transcriptional activators as well. In particular, AREB6 has

been shown to bind to the promoter of the housekeeping gene Na,K- ATPase α-1 and to repress gene expression dependent on cell type and on the context of the binding site (Watanabe *et al.*, 1993, J.Biochem.,114, p. 849-855).

SIP1 mRNA expression in mice

Northern analysis demonstrated the presence of a major SIP1 6 kb mRNA in the embryo and several tissues of adult mice, with very weak expression in liver and testis. A minor 9 kb-long transcript is also detected, which is however present in the 7 dpc embryo. *In situ* hybridization documented SIP1 transcription in the 7.5 dpc embryo in the extraembryonic and embryonic mesoderm. The gene is weakly expressed in embryonic ectoderm. In the 8.5 dpc embryo, very strong expression is seen in extraembryonic mesoderm (blood islands), neuroepithelium and neural tube, the first and second branchial arches, the optic eminence, and predominantly posterior presomitic mesoderm. Weaker but significant expression is detected in somites and notochord. Between day 8.5 and 9.5, this pattern extends clearly to the trigeminal and facio-acoustic neural crest tissue. Around midgestation, the SIP1 gene is expressed in the dorsal root ganglia, spinal cord, trigeminal ganglion, the ventricular zone of the frontal cortex, kidney mesenchyme, non-eptihelial cells of duodenum and midgut, pancreatic primordium, urogenital ridge and gonads, the lower jaw and the snout region, cartilage primordium in the humerus region, the primordium of the clavicle and the segmental precartilage sclerotome-derived condensations along the vertebral axis. SIP1 mRNA can also be detected in the palatal shelf, lung mesenchyme, stomach and inferior ganglion of vagus nerve. In addition, primer extension analysis has demonstrated the presence of SIP1 mRNA in embryonic stem cells. It is striking that the expression of SIP1 in the 8.5 dpc embryo in the blood islands and presomitic mesoderm coincides with tissues affected in BMP-4 knockout mice, which have been shown to die between 6.5 and 9.5 dpc with a variable phenotype. These surviving till later stages of development showed disorganized posterior structures and a reduction in extraembryonic mesoderm, including blood islands (Winnier *et al.*, 1995, Genes Dev.,9, 2105-2116).

The mRNA expression of δ-EF1 proteins has been documented as well. In mouse, δ-EF1 mRNA has been detected in mesodermal tissues such as notochord,

somites and nephrotomes, and in other sites such as the nervous sytem and the lens in the embryo (Funahashi *et al.*, 1993, Development, 119, p.433-446). In adult hamster, δ-EF1 mRNA has been detected in the cells of the endocrine pancreas, anterior pituitary and central nervous system (Franklin *et al.*, 1994, Mol.Cell.Biol.,14, p. 6773-6788). The majority of these δ-EF1 and SIP1 expression sites overlap with sites where the restricted expression pattern of certain type I STK receptors (such as ALK-4/ActR-IA, and ALK-6/BMPR-IB) has been documented (Verschueren *et al.*, 1995, Mech.Dev.,52, p.109-123).

SIP2

Characterization of SIP2

SIP2 was picked up initially as a two hybrid clone of 1052 bp (th12) that shows interaction in yeast with Smad1, 2 and 5 C-terminal domains and full-size Smad1. Using GST-pull down experiments (as described for SIP1) also an interaction with Smad1, 2 and 5 C-terminal domains *in vitro* have been demonstrated.

a) SIP2 full length sequence

Th12 showed high homology to a partial cDNA (KIAA0150) isolated from the human myoloblast cell line KG1. However, this human cDNA is +/- 2 kb longer at the 3' end of th12. Using this human cDNA, an EST library was screened and mouse EST were detected homologous to the 3'end of KIAA0150 cDNA. Primers were designed based on th12 sequence and the mouse EST found to amplify a cDNA that contains the stop codon at the 3'end.

5' sequences encompassing the start codon was obtained using 5'RACE-PCR .

Gene bank accession numbers for the mentioned EST clones used to complete the SIP2 open reading frame:

Human KIAA0150 ; D63484

Mouse EST sequence; Soares mouse p3NMF19.5; W82188,

Primers used to reconstitute SIP2 open reading frame:

based on th12 sequence: F3th12F (forward primer) 5'-cggcggcagatacgcctcctgca

based on EST sequence: th12mouse1 (reverse primer) 5'-caggagcagttgtgggtagagccttcatc

Primers used for 5'-race;

all are reverse primers derived from th12 sequence

1: 5'-ctggactgagctggacctgtctctccagtac

2 : 5'-cacaagggagtatttcttgcgccacgaagg

3: 5'-gccatggtgtgaggagaagc

The full size SIP2 deduced from the assembly of these sequences contains 950 amino acids as depicted in SEQ ID NO.4, while the nucleotide sequence is depicted in SEQ.ID.NO.3.

b) SIP2 sequence homologies

SIP2 contains a domain encompassing 5 CCCH type zinc fingers. This domain was found in other protein such as Clipper in Drosophila, No Arches in Zebrafish and CPSF in mammals. No Arches is essential for development of the branchial arches in Zebrafish and CPSF is involved in trancription termination and polyadenylation. The domain containing the 5 CCCH in Clipper was shown to have an EndoRNase activity (see below).

c) SIP2 CCCH domain has an RNAse activity

The domain containing the 5 CCCH -type zinc fingers of SIP2 was fused to GST and the fusion protein was purified from E.coli. This fusion protein displays a RNAse activity when incubated with labeled RNA produced *in vitro*. In addition, it has been ~~shown that this fusion protein was able to bind single stranded DNA.~~

In more detail :

GST fusion proteins of SIP2 5xCCCH; PLAG1 (an unrelated zinc finger protein), $SIP1_{CZF}$ (C-terminal zinc finger cluster of SIP1) and th1 (SIP1 partial polypeptide

isolated in the two-hybrid screening), and cytoplasmatic tail of CD40 were produced in E.coli and purified using glutathione sepharose beads. Three ^{35}S labeled substrates, previously used to demonstrate the RNAse activity of Clipper, a related protein from Drosophila (Bai, C. and Tolias P.P. 1996, cleavage of RNA Hairpins Mediated by a Developmentally Regulated CCCH Zinc Finger Protein. Mol Cell. Biol. 16: 6661-6667) were produced by *in vitro* transcription. The RNA cleavage reactions with purified GST fusion proteins were performed in the presence of RNAsin (blocking RNAseA activity). Equal aliquots of each reaction were taken out at time points 1', 7', 15', 30', 60'. Degradation productes were separated on a denaturing polyacrylamide gel and visualized by autoradiography. These experiments demonstrated that GST-SIP2 5XCCCH has an RNAse activity and degrades all tested substrates, while GST-PLAG1, GST-CD40, GST-SIP1$_{CZF}$ and GST-th1 do not have this activity.

d) Interaction between th12 (partial SIP2 polypeptide) and Smad C-domains in GST pull down experiments.

C-domains of *Xenopus (X)*Smad1 and mouse Smad2 and 5 were produced in E. coli as fusion proteins with gluthatione S-transferase and coupled to gluthatione beads. An unrelated GST-fusion protein (GST-CD40 cytoplasmatic mail) and GST itself were used as negative controls.

Th12 protein, provided with an HA-tag at its N-terminal end, was produced in Hela cells using the T7 vaccinia virus expression system and metabolically labeled. Expression of Th12 was confirmed by immune precipitation with HA antibody, followed by SDS-page and autoradiography. Th12 protein is produced as a ± 50 kd protein. Cell extracts prepared from Hela cells expressing this protein were mixed with GST-Smad C-domain beads in GST pull down buffer and incubated overnight at 4° C. The beads were then washed four times in the same buffer, the bound proteins eluted in Laemmli sample buffer and separated by SDS-PAGE. "Pulled down" th12 protein was visualized by Western blotting , using HA antibody. These experiments demonstrate that th12 is efficiently pulled down by GST-Smad C-domain beads, and not by GST-CD40 or GST alone.

Conclusion on SIP2

SIP2 is a Smad interacting protein that contains a RNAse activity. The finding that Smads interact with potential RNAses provides an unexpected link between the TGF-β signal transduction and mRNA stabilisation.

SIP5

Characterization of SIP5

One contiguous open reading frame is fused in frame to the GAL4 transactivating domain in the two hybrid vector pACT-2 (Clontech). This represents a partial cDNA, since no in frame translational stop codon is present. The sequence has no significant homology to anything in the database, but displays a region of high homology with following EST clones:

Mouse: accession numbers: AA212269 (Stratagene mouse melanom); AA215020 (Stratagene mouse melanom), AA794832 (Knowles Solter mouse 2 c) and Human: accession numbers AA830033, AA827054, AA687275, AA505145, AA371063.

Analysis of interaction of the SIP5 prey protein with different bait proteins (which are described in the data section obtained with SIP1) in a yeast two hybrid assay can be summarized as follows

Empty bait vector pGBT9	-
Full length XSmad1	+
Xsmad1 C-domain	+
Xsmad1 C-domain with G418S substitution	+
Mouse Smad2 C-domain	+
Mouse Smad5 C-domain	+
Lamin (pLAM; Clontech)	-

SIP5 partial protein encoded by above described cDNA also interacts with Xsmad1, mouse Smad2 and 5 C-domains in vitro as analysed by the GST pull down assay (previously described for SIP1 and SIP2). Briefly, the partial SIP5 protein was tagged with a myc tag at its C-terminal end and expressed in COS-1 cells. GST-Smad C-domain fusion proteins, GST-CD40 cytoplasmatic tail and GST alone were expressed in E. coli and coupled to glutathione sepharose beads. These beads were subsequently used to pull down partial SIP5 protein from COS cell lysates, as was demonstrated after SDS-PAGE of pulled down proteins followed by Western blotting using anti myc antibody. In this assay, SIP5 was pulled down by GST-Xsmad1, 2 and 5 C-domains, but not by GSTalone or GST-CD40.
A partial, but coding, nucleic acid sequence for SIP5 is depicted in SEQ.ID.NO.10.

SIP7

Characterization of SIP7

One contiguous open reading frame is fused in frame to the GAL4 transactivating domain in the two hybrid vector pACT2. This is a partial clone, since no in frame translational stop codon is present. Part of this clone shows homology to Wnt-7b,accession number: M89802, but the clone seems to be a novel cDNA or a cloning artefact. The homology of the SIP7 cDNA with the known Wnt7-b cDNA starts at nucleotide 390 and extends to nucleotide 846. This corresponds to the nucleotides 74-530 in Wnt7-b coding sequences (with A of the translational start codon considered as nucleotide nr 1). In SIP7 cDNA this region of homology is preceded by a sequence that shows no homology to anything in the database. It is not clear whether the SIP7 cDNA is for example a new Wnt7-b transcript or whether it is a scrambled clone as a result of the fusion of two cDNAs during generation of the cDNA library.

Analysis of the interaction of the SIP7 prey protein with different bait proteins in a yeast two hybrid assay can be summarized as follows:

PGBT9	-
Full length XSmad1	-
Xsmad1 C-domain	+
Xsmad1 C-domain, G418S	+
Xsmad1 C-domain del aa 424-466	-
Xsmad1 N-terminal domain	-
Mouse smad2 C-domain	+
Mouse Smad5 C-domain	+
Lamin (pLAM)	-

SIP7 partial protein encoded by above described cDNA also interacts with Xsmad1, mouse Smad2 and 5 C-domains in vitro as analysed by the GST pull down assay, as descibed above for SIP5. In this assay, N-terminally myc-tagged SIP7 protein was specifically pulled down by GST-Xsmad1, 2 and 5 C-domains, but not by GSTalone or GST-CD40.

A partial, but coding, nucleic acid sequence for SIP7 is depicted in SEQ.ID.NO.8.

General description of the methods used

Plasmids and DNA manipulations

Mouse Smad1 and Smad2 cDNAs used in this study were identified by low stringency screening of oligo-dT primed λEx*lox* cDNA library made from 12 dpc mouse embryos (Novagen), using Smad5 (MLP1.2 clone as described in Meersseman et al., 1997, Mech.Dev.,61, p.127-140) as a probe. The same library was used to screen for full-size SIP1, and yielded λExTW6. The tw6 cDNA was 3.6 kb long, and overlapped with th1 cDNA, but contained additional 3'-coding sequences including an in-frame stop codon. Additional 5' sequences were obtained by 5' RACE using the Gibco-BRL 5' RACE kit.

XSmad1 full-size and C-domain bait plasmids were constructed using previously described *Eco*RI-*Xho*I inserts(Meersseman et al.,1997, Mech.Dev.,61, p.127-140), and cloned between the *Eco*RI and *Sal*I sites of the bait vector pGBT-9 (Clontech), such that in-frame fusions with $GAL4_{DBD}$ were obtained. Similar bait

plasmids with mouse Smad1, Smad2 and Smad5 were generated by amplifying the respective cDNA fragments encoding the C-domain using Pfu polymerase (Stratagene) and primers with *Eco*RI and *Xho*I sites. The G418S *X*Smad1 C-domain was generated by oligonucleotide-directed mutagenesis (Biorad).

To generate in-frame fusions of Smad C-domains with GST, the same Smad fragments were cloned in pGEX-5X-1 (Pharmacia). The phage T7 promoter-based SIP1 (TH1) construct for use in the T7VV system was generated by partial restriction of the th1 prey cDNA with *Bgl*II, followed by restriction with *Sal*I, such that SIP1 (TH1) was lifted out of the prey vector along with an in-frame translational start codon, an HA-epitope tag of the flu virus, and a stop codon. This fragment was cloned into pGEM-3Z (Promega) for use in the T7VV system. A similar strategy was used to clone SIP2 (th12) into pGEM-3Z.

PolyA$^+$ RNA from 12.5 dpc mouse embryos was obtained with oligotex-dT (Qiagen). Randomly primed cDNA was prepared using the Superscript Choice system (Gibco-BRL). cDNA was ligated to an excess of Sfi double-stranded adaptors containing *Stu*I and *Bam*HI sites. To facilitate cloning of the cDNAs, the prey plasmid pAct (Clontech) was modified to generate pAct/Sfi-Sfi. Restriction of this plasmid with *Sfi* generates sticky ends which are not complementary, such that self-ligation of the vector is prevented upon cDNA cloning. A library containing 3.6×10^6 independent recombinant clones with an average insert size of 1,100 bp was obtained.

Synthesis of SIP1 and GST pull-down experiments

Expression of SIP1 (TH1) and SIP2 (TH12) in mammalian cells with the T7VV system and the preparation of the cell lysates were as described previously (Verschueren, K et al.,1995, Mech.Dev.,52, p.109-123).

GST fusion proteins were expressed in *E. coli* (strain BL21) and purified on gluthathione-Sepharose beads (Pharmacia). The beads were washed first four times with PBS supplemented with protease inhibitors, and then mixed with 50 µl of lysate (prepapred from T7VV-infected SIP1-expressing mammalian cells) in 1 ml of GST buffer (50 mM Tris-HCl pH 7.5, 120 mM NaCl, 2 mM EDTA, 0.1% (v/v) NP-40, and protease inhibitors). They were mixed at 4°C for 16 hours. Unbound proteins were

removed by washing the beads four times with GST buffer. Bound proteins were harvested by boiling in sample buffer, and resolved by SDS-PAGE. Separated proteins were visualized using autoradiography or immunodetection after Western blotting; using anti-HA monoclonal antbody (12CA5) and alkaline phosphatase-conjugated anti-mouse 2ary antibody (Amersham).

EMSA(=electrophoretic mobility shift assay)

The sequence of the κE2 WT and mutated κE2 oligonucleotides are identical as disclosed in Sekido et al; (1994, Mol.Cell.Biol.,14, p. 5692-5700). The sequence of the AREB6 oligonucleotide was obtained from Ikeda et al;(1995, Eur.J.Biochem, 233, p. 73-82). IL2 oligonucleotide is depicted in Williams et al;(1991, Science, 254, p.1791-1794).

The sequence of Brachyury binding site is 5'-TGACACCTAGGTGTGAATT-3'. The negative control GATA2 oligonucleotide sequences originated from the endothelin promoter (Dorfman et al; 1992, J.Biol.Chem., 267, p. 1279-1285). Double stranded oligonucleotides were labeled with polynucleotide kinase and ^{32}P γ-ATP and purified from a 15% polyacrylamide gel. Gel retardation assays were performed according to Sekido et al; (1994, Mol.Cell.Biol.,14, p. 5692-5700).

RESULTS OF TWO HYBRID SCREENING (Xsmad1 C-domain bait versus 12.5 dpc mouse embryo library; 600.000 recombinant clones screened in 4x 10^6 yeasts).

SIP 1 - Three independent clones isolated (th1, th88 and th94)

- Zinc-finger-homeodomain protein
- Homology to δEF-1 (see above)
- Interactions in yeast:

XSmad1 C-domain bait	+
Empty bait	-
Lamin	-

XSmad1 full length	-
XSmad1 N-domain	-
mSmad1 C-domain	+
mSmad2 C-domain	+
mSmad5 C-domain	+
XSmad1 C-domain del 424-466	-
XSmad1 C-domain G418S	+

* Interaction with C-domain of XSmad1 and mSmads confirmed in vitro using GST-pull downs and co-immunoprecipitations
* Extended clone (TW6) isolated through library screening using th1 sequences as a probe
* C-terminal TW6 zinc-finger cluster binds to E2 box sequences (cfr δEF-1), Brachyury T binding site, Brachyury promoter sequences

SIP2 also called clone TH12- Three independent clones isolated (th12,th73,th93)

Highly homologous to KIAA0150 gene product, isolated from the myeloblast cell line KG1(Ref: Nagase et al. 1995; DNA Res 2 (4) 167-174.

- Interactions in yeast:

XSmad1 C-domain bait	+
Empty bait	-
Lamin	-
XSmad1 full length	+
XSmad1 N-domain	ND
mSmad1 C-domain	+
mSmad2 C-domain	+
mSmad5 C-domain	+
XSmad1 C-domain del 424-466	-
XSmad1 C-domain G418S	+

TH60 - Two independent clones isolated (th60 and th77)

- Zinc finger protein

  homology to snail (transcriptional repressor) and to ATBF1 (complex homeodomain zinc finger protein)

- Interactions in yeast:

XSmad1 C-domain bait	+
Empty bait	-
Lamin	-

TH72 - One clone isolated

- Encodes a partial DPC-4 (Smad4) cDNA (see above)
- Interactions in yeast:

XSmad1 C-domain bait	+++
Empty bait	-
Lamin	-
XSmad1 full length	ND
XSmad1 N-domain	-
mSmad1 C-domain	+++
mSmad2 C-domain	ND
mSmad5 C-domain	+++
XSmad1 C-domain del 424-466	-
XSmad1 C-domain G418S	+

SIP5 (also called clone th76).

Analysis of interaction of the SIP5 prey protein with different bait proteins (which are described in the data section obtained with SIP1) in a yeast two hybrid assay can be summarized as follows

Empty bait vector pGBT9	-
~~Full length XSmad1~~	~~+~~
Xsmad1 C-domain	+
Xsmad1 C-domain G418S	+
Mouse Smad2 C-domain	+

Mouse Smad5 C-domain	+
Lamin (pLAM; Clontech)	-

SIP7 (also called clone th74)

Analysis of the interaction of the SIP7 prey protein with different bait proteins in a yeast two hybrid assay can be summarized as follows:

PGBT9	-
Full length XSmad1	-
Xsmad1 C-domain	+
Xsmad1 C-domain, G418S	+
Xsmad1 C-domain del aa 424-466	-
Xsmad1 N-terminal domain	-
Mouse smad2 C-domain	+
Mouse Smad5 C-domain	+
Lamin (pLAM)	-

The following clones have been investigated less extensively. They are considered as "true positives" because they interact with the XSmad1 C-domain bait and not with the empty bait (i.e GAL-4 DBD alone)

TH75: -Three independent clones isolated (th75, th83, th89)

-Partial aa sequences do not show significant homology to proteins in the public databases

- Interactions in yeast:

XSmad1 C-domain bait	+++
Empty bait	-

TH92: -Zinc finger protein

-homology to KUP

TH79, TH86, TH90, : Partial sequences do not display significant homology to any protein sequence in the public databases.

<u>**Clones available in the sequence listing as conversion table from clone notation to sequence listing notation**</u>

SIP 1 nucleotide sequence	= SEQ ID NO 1
SIP 1 amino acid sequence	= SEQ ID NO 2
SIP 2 nucleotide sequence	= SEQ ID NO 3
SIP 2 amino acid sequence	= SEQ ID NO 4
TH60(TH77)	= SEQ ID NO 5
TH72 (DPC4 or Smad4)	= SEQ ID NO 6
TH72\R	= SEQ ID NO 7
SIP 7(th74)	= SEQ ID NO 8
TH75F(TH83F,TH89F)	= SEQ ID NO 9
SIP 5(th76)	= SEQ ID NO 10
TH79F	= SEQ ID NO 11
TH79R	= SEQ ID NO 12
TH83R	= SEQ ID NO 13
TH86F	= SEQ ID NO 14
TH86R	= SEQ ID NO 15
TH89=TH75R	= SEQ ID NO 16
TH90F	= SEQ ID NO 17
TH90R	= SEQ ID NO 18
TH92F	= SEQ ID NO 19
TH92R	= SEQ ID NO 20

LEGEND TO FIGURE 1

XSmad1 C-domain interacts with SIP1 in mammalian cells and deletion of the 51 aa long SBD (Smad binding domain) in SIP1 abolishes the interaction.

COS-1 cells were transiently transfected with expression constructs encoding N-terminally myc-tagged SIP1 and a GST-XSmad1 C-domain fusion protein. The latter was purified from cell extracts using gluthatione-sepharose beads. Purified proteins were visualized after SDS-PAGE and Western blotting using anti-GST antibody (Pharmacia), (Panel A, slim arrow).

Myc-tagged SIP1 protein was co-purified with GST-XSmad1 C-domain fusion protein, as was shown by Western blotting of the same material using anti-myc monoclonal antibody (Santa Cruz)(Panel C, lane one, fat arrow). Deletion of the 51 aa long SBD in SIP1 abolished this interaction (panel C, lane 2). Note that the amounts of purified GST-XSmad1 C-domain protein and levels of expression of both SIP1 (wild type and SIP1del SBD) proteins in total cell extracts were comparable (compare lanes 1 and 2 in panel A and B).

SEQUENCE LISTING

SEQ ID NO 1

1	GCAGCACTCA	GCACCAAATG	CTAACCCAAG	GAGCAGGTAA	CCGCAAGTTC	AAGTGCACGG
61	AGTGTGGCAA	GGCCTTCAAG	TACAAGCACC	ACCTGAAAGA	ACACCTGAGA	ATTCACAGTG
121	GTGAAAAACC	TTACGAATGC	CCAAACTGCA	AGAAACGCTT	CTCTCATTCT	GGGTCCTACA
181	GTTCACATAT	CAGCAGCAAG	AAATGTATTG	GTTTAATATC	AGTAAATGGC	CGAATGAGAA
241	ACAATATCAA	GACGGGTTCT	TCCCCTAATT	CTGTTTCTTC	TTCTCCTACT	AACTCAGCCA
301	TTACTCAGTT	AAGGAACAAG	TTGGAAAATG	GAAAACCACT	TAGCATGTCT	GAGCAGACAG
361	GCTTACTTAA	GATTAAAACA	GAACCACTAG	ACTTCAATGA	CTATAAAGTT	CTTATGGCAA
421	CACATGGGTT	TAGTGGCAGC	AGTCCCTTTA	TGAACGGTGG	GCTTGGAGCC	ACCAGCCCTT
481	TAGGTGTACA	CCCATCTGCT	CAGAGTCCAA	TGCAGCACTT	AGGTGTAGGG	ATGGAAGCCC
541	CTTTACTTGG	ATTTCCCACT	ATGAATAGTA	ACTTGAGTGA	GGTACAAAAG	GTTCTACAGA
601	TTGTGGACAA	TACGGTTTCT	AGGCAAAAGA	TGGACTGCAA	GACGGAAGAC	ATTTCAAAGT
661	TGAAAGGTTA	TCACATGAAG	GATCCATGTT	CTCAGCCAGA	AGAACAAGGG	GTAACTTCTC
721	CCAATATTCC	CCCTGTCGGT	CTTCCAGTAG	TGAGTCATAA	CGGTGCCACT	AAAAGTATTA
781	TTGACTATAC	CTTAGAGAAA	GTCAATGAAG	CCAAAGCTTG	CCTCCAGAGC	TTGACCACCG
841	ACTCAAGGAG	ACAGATCAGT	AACATAAAGA	AAGAGAAGTT	GCGTACTTTG	ATAGATTTGG
901	TCACTGATGA	TAAAATGATT	GAGAACCACA	GCATATCCAC	TCCATTTTCA	TGCCAGTTCT
961	GTAAAGAAAG	CTTCCCGGGC	CCTATTCCCC	TGCATCAGCA	TGAACGATAC	CTGTGTAAGA
1021	TGAATGAAGA	GATCAAGGCA	GTCCTGCAAC	CTCATGAAAA	CATAGTCCCC	AACAAAGCTG
1081	GAGTTTTTGT	TGATAATAAA	GCCCTCCTCT	TGTCATCTGT	ACTTTCCGAG	AAAGGACTGA
1141	CAAGCCCCAT	CAACCCATAC	AAGGACCACA	TGTCTGTACT	GAAAGCATAC	TATGCTATGA
1201	ACATGGAGCC	CAACTCTGAT	GAACTGCTGA	AAATCTCCAT	TGCTGTGGGC	CTTCCTCAGG
1261	AATTTGTGAA	GGAATGGTTT	GAGCAAAGAA	AAGTCTACCA	GTATTCGAAT	TCCAGGTCAC
1321	CATCACTGGA	AAGGACCTCC	AAGCCGTTAG	CTCCCAACAG	TAACCCCACC	ACAAAAGACT
1381	CTTTGTTACC	CAGGTCTCCT	GTAAAACCTA	TGGACTCCAT	CACATCGCCA	TCTATAGCAG
1441	AACTCCACAA	CAGTGTTACG	AGTTGTGATC	CTCCTCTCAG	GCTAACAAAA	TCTTCCCATT
1501	TCACCAATAT	TAAAGCAGTT	GATAAACTGG	ACCACTCGAG	GAGTAATACT	CCTTCTCCTT
1561	TAAATCTTTC	CTCCACATCT	TCTAAAAACT	CCCACAGTAG	CTCGTACACT	CCAAATAGCT
1621	TCTCTTCCGA	GGAGCTGCAG	GCTGAGCCGT	TGGACCTGTC	ATTACCAAAA	CAAATGAGAG
1681	AACCCAAAGG	TATTATAGCC	ACAAAGAACA	AAACAAAAGC	TACTAGCATA	AACTTAGACC
1741	ACAACAGTGT	TTCTTCATCG	TCTGAGAATT	CAGATGAGCC	TCTGAATTTG	ACTTTTATCA
1801	AGAAAGAGTT	TTCAAATTCT	AATAACCTGG	ACAATAAAAG	CAACAACCCT	GTGTTCGGCA
1861	TGAACCCATT	TAGTGCCAAG	CCTTTATACA	CCCCTCTTCC	ACCACAGAGC	GCATTTCCCC
1921	CTGCCACTTT	CATGCCACCA	GTCCAGACCA	GCATCCCCGG	GCTACGACCA	TACCCAGGAC
1981	TGGATCAGAT	GAGCTTCCTA	CCGCATATGG	CCTATACCTA	CCCAACGGGA	GCAGCTACCT
2041	TTGCTGATAT	GCAGCAAAGG	AGGAAATACC	AGAGGAAACA	AGGATTTCAG	GGAGACTTGC
2101	TGGATGGAGC	ACAAGACTAC	ATGTCAGGCC	TAGATGACAT	GACAGACTCC	GATTCCTGTC
2161	TGTCTCGAAA	GAAGATAAAG	AAGACAGAAA	GTGGCATGTA	TGCATGTGAC	TTATGTGACA
2221	AGACATTCCA	GAAAAGCAGT	TCCCTTCTGC	GACATAAATA	CGAACACACA	GGAAAGAGAC

2281	CACACCAGTG	TCAGATTTGT	AAGAAAGCGT	TCAAACACAA	ACACCACCTT	ATCGAGCACT
2341	CGAGGCTGCA	CTCGGGCGAG	AAGCCCTATC	AGTGTGACAA	ATGTGGCAAG	CGCTTCTCAC
2401	ACTCGGGCTC	CTACTCGCAG	CACATGAATC	ACAGGTACTC	CTACTGCAAG	CGGGAGGCGG
2461	AGGAGCGGGA	AGCAGCCGAG	CGCGAGGCGC	GAGAGAAAGG	GCACTTGGGA	CCCACCGAGC
2521	TGCTGATGAA	CCGGGCTTAC	CTGCAGAGCA	TCACCCCTCA	GGGGTACTCT	GACTCGGAGG
2581	AGAGGGAGAG	CATGCCGAGG	GATGGCGAGA	GCGAGAAGGA	GCACGAGAAG	GAGGGCGAGG
2641	AGGGTTATGG	GAAGCTGCGG	AGAAGGGACG	GCGACGAGGA	GGAAGAGGAG	GAAGAGGAAG
2701	AAAGTGAAAA	TAAAAGTATG	GATACGGATC	CCGAAACGAT	ACGGGATGAG	GAAGAGACTG
2761	GGGATCACTC	GATGGACGAC	AGTTCAGAGG	ATGGAAAAAT	GGAAACCAAA	TCAGACCACG
2821	AGGAAGACAA	TATGGAAGAT	GGCATGGGAT	AAACTACTGC	ATTTTAAGCT	TCCTATTTTT
2881	TTTTTCCAGT	AGTATTGTTA	CCTGCTTGAA	AACACTGCTG	TGTTAAGCTG	TTCATGCACG
2941	TGCCTGACGC	TTCCAGGAAG	CTGTAGAGAG	GGACAAAAAG	GGGCACTTCA	GCCAAGTCTG
3001	AGTTAG					

SEQ ID NO 2

1	MetLeuThrGlnGly	AlaGlyAsnArgLys	PheLysCysThrGlu
16	CysGlyLysAlaPhe	LysTyrLysHisHis	LeuLysGluHisLeu
31	ArgIleHisSerGly	GluLysProTyrGlu	CysProAsnCysLys
46	LysArgPheSerHis	SerGlySerTyrSer	SerHisIleSerSer
61	LysLysCysIleGly	LeuIleSerValAsn	GlyArgMetArgAsn
76	AsnIleLysThrGly	SerSerProAsnSer	ValSerSerSerPro
91	ThrAsnSerAlaIle	ThrGlnLeuArgAsn	LysLeuGluAsnGly
106	LysProLeuSerMet	SerGluGlnThrGly	LeuLeuLysIleLys
121	ThrGluProLeuAsp	PheAsnAspTyrLys	ValLeuMetAlaThr
136	HisGlyPheSerGly	SerSerProPheMet	AsnGlyGlyLeuGly
151	AlaThrSerProLeu	GlyValHisProSer	AlaGlnSerProMet
166	GlnHisLeuGlyVal	GlyMetGluAlaPro	LeuLeuGlyPhePro
181	ThrMetAsnSerAsn	LeuSerGluValGln	LysValLeuGlnIle
196	ValAspAsnThrVal	SerArgGlnLysMet	AspCysLysThrGlu
211	AspIleSerLysLeu	LysGlyTyrHisMet	LysAspProCysSer
226	GlnProGluGluGln	GlyValThrSerPro	AsnIleProProVal
241	GlyLeuProValVal	SerHisAsnGlyAla	ThrLysSerIleIle
256	AspTyrThrLeuGlu	LysValAsnGluAla	LysAlaCysLeuGln
271	SerLeuThrThrAsp	SerArgArgGlnIle	SerAsnIleLysLys
286	GluLysLeuArgThr	LeuIleAspLeuVal	ThrAspAspLysMet
301	IleGluAsnHisSer	IleSerThrProPhe	SerCysGlnPheCys

316	LysGluSerPhePro	GlyProIleProLeu	HisGlnHisGluArg
331	TyrLeuCysLysMet	AsnGluGluIleLys	AlaValLeuGlnPro
~~346~~	~~HisGluAsnIleVal~~	~~ProAsnLysAlaGly~~	~~ValPheValAspAsn~~
361	LysAlaLeuLeuLeu	SerSerValLeuSer	GluLysGlyLeuThr
376	SerProIleAsnPro	TyrLysAspHisMet	SerValLeuLysAla
391	TyrTyrAlaMetAsn	MetGluProAsnSer	AspGluLeuLeuLys
406	IleSerIleAlaVal	GlyLeuProGlnGlu	PheValLysGluTrp
421	PheGluGlnArgLys	ValTyrGlnTyrSer	AsnSerArgSerPro
436	SerLeuGluArgThr	SerLysProLeuAla	ProAsnSerAsnPro
451	ThrThrLysAspSer	LeuLeuProArgSer	ProValLysProMet
466	AspSerIleThrSer	ProSerIleAlaGlu	LeuHisAsnSerVal
481	ThrSerCysAspPro	ProLeuArgLeuThr	LysSerSerHisPhe
496	ThrAsnIleLysAla	ValAspLysLeuAsp	HisSerArgSerAsn
511	ThrProSerProLeu	AsnLeuSerSerThr	SerSerLysAsnSer
526	HisSerSerSerTyr	ThrProAsnSerPhe	SerSerGluGluLeu
541	GlnAlaGluProLeu	AspLeuSerLeuPro	LysGlnMetArgGlu
556	ProLysGlyIleIle	AlaThrLysAsnLys	ThrLysAlaThrSer
571	IleAsnLeuAspHis	AsnSerValSerSer	SerSerGluAsnSer
586	AspGluProLeuAsn	LeuThrPheIleLys	LysGluPheSerAsn
601	SerAsnAsnLeuAsp	AsnLysSerAsnAsn	ProValPheGlyMet
616	AsnProPheSerAla	LysProLeuTyrThr	ProLeuProProGln
631	SerAlaPheProPro	AlaThrPheMetPro	ProValGlnThrSer
646	IleProGlyLeuArg	ProTyrProGlyLeu	AspGlnMetSerPhe
661	LeuProHisMetAla	TyrThrTyrProThr	GlyAlaAlaThrPhe
676	AlaAspMetGlnGln	ArgArgLysTyrGln	ArgLysGlnGlyPhe
691	GlnGlyAspLeuLeu	AspGlyAlaGlnAsp	TyrMetSerGlyLeu
706	AspAspMetThrAsp	SerAspSerCysLeu	SerArgLysLysIle
721	LysLysThrGluSer	GlyMetTyrAlaCys	AspLeuCysAspLys
736	ThrPheGlnLysSer	SerSerLeuLeuArg	HisLysTyrGluHis
751	ThrGlyLysArgPro	HisGlnCysGlnIle	CysLysLysAlaPhe
766	LysHisLysHisHis	LeuIleGluHisSer	ArgLeuHisSerGly
781	GluLysProTyrGln	CysAspLysCysGly	LysArgPheSerHis
796	SerGlySerTyrSer	GlnHisMetAsnHis	ArgTyrSerTyrCys
811	LysArgGluAlaGlu	GluArgGluAlaAla	GluArgGluAlaArg
826	GluLysGlyHisLeu	GlyProThrGluLeu	LeuMetAsnArgAla
841	TyrLeuGlnSerIle	ThrProGlnGlyTyr	SerAspSerGluGlu

856 ArgGluSerMetPro ArgAspGlyGluSer GluLysGluHisGlu
871 LysGluGlyGluGlu GlyTyrGlyLysLeu ArgArgArgAspGly
886 AspGluGluGluGlu GluGluGluGluGlu SerGluAsnLysSer
901 MetAspThrAspPro GluThrIleArgAsp GluGluGluThrGly
916 AspHisSerMetAsp AspSerSerGluAsp GlyLysMetGluThr
931 LysSerAspHisGlu GluAspAsnMetGlu AspGlyMetGly

SEQ ID NO 3

```
   1 CTGGCTAGGC GTCGCGGACT CCGGAGATGG AGGAAAAGGA GCAGCTGCGG CGGCAGATAC
  61 GCCTCCTGCA GGGTCTAATT GATGACTATA AAACACTCCA CGGCAATGGC CCTGCCCTGG
 121 GCAACTCATC AGCTACTCGG TGGCAGCCAC CCGTGTTCCC GGGTGGCAGG ACCTTTGGCG
 181 CCCGCTACTC CCGTCCAAGT CGGAGGGGCT TCTCCTCACA CCATGGCCCT TCGTGGCGCA
 241 AGAAATACTC CCTTGTGAAT CAGCCTGTGG AATCTTCTGA CCCAGCCAGC GATCCTGCTT
 301 TTCAGACATC CCTCAGGTCT GAGGATAGCC AGCATCCTGA ACCCCAGCAG TATGTACTGG
 361 AGAGACAGGT CCAGCTCAGT CCAGATCAGA ATATGGTTAT TAAGATCAAG CCACCATCAA
 421 AGTCAGGTGC CATCAATGCT TCAGGGGTCC AGCGGGGGTC CTTGGAAGGC TGTGATGACC
 481 CCTCTTGGAG TGGCCAAAGA CCCCAAGGAA GTGAGGTTGA GGTCCCTGGT GGACAACTGC
 541 AGCCTGCAAG GCCAGGAAGA ACCAAGGTGG GTTACAGTGT GGACGACCCC CTCTTGGTCT
 601 GCCAGAAGGA GCCTGGCAAG CCTCGGGTAG TGAAGTCTGT GGGCAGGGTG AGTGACAGCT
 661 CTCCCGAGCA TCGGCGGACA GTCAGTGAAA ATGAAGTGGC CCTCAGGGTA CACTTCCCAT
 721 CTGTCCTGCC CCATCACACT GCTGTGGCTC TGGGCAGGAA GGTAGGCCCT CATTCTACCA
 781 GCTATTCTGA ACAGTTCATT GGAGACCAAA GAGCAAACAC TGGCCACTCA GACCAGCCAG
 841 CTTCCTTGGG GCCAGTGGTG GCTTCAGTCA GACCAGCAAC AGCCAGGCAG GTCAGGGAGG
 901 CCTCACTGCT CGTGTCCTGT CGAACCAGCA AGTTTCGGAA AAACAACTAC AAATGGGTAG
 961 CTGCCTCAGA AAAGAGCCCA CGGGTCGCTC GGAGAGCCCT CAGTCCCAGA ACAACTCTGG
1021 AGAGCGGGAA CAAGGCCACT TTGGGTACAG TTGGAAAGAC AGAGAAGCCA CAGCCTAAAG
1081 TTGACCCAGA GGTGAGGCCG GAGAAACTGG CCACACCATC CAAGCCTGGC CTCTCTCCCA
1141 GCAAGTACAA GTGGAAGGCT TCCAGCCCGT CTGCTTCCTC CTCTTCCTCT TTCCGTTGGC
1201 AGTCTGAGGC TGGCAGCAAG GACCATACTT CTCAGCTCTC CCCAGTCCCA TCTAGGCCCA
1261 CATCAGGGGA CAGACCAGCA GGGGGACCCA GCAGCTTGAA GCCCCTCTTT GGAGAGTCAC
1321 AGCTCTCAGC TTACAAAGTG AAGAGCCGGA CCAAGATTAT CCGGAGGCGG GGCAATACCA
1381 GCATTCCTGG GGACAAGAAG AACAGCCCTA CAACTGCCAC CACCAGCAAA AACCATCTTA
1441 CCCAGCGACG GAGACAGGCC CTCCGGGGGA AGAATAGCCC GGTTCTAAGG AAGACTCCCC
1501 ACAAGGGTCT GATGCAGGTC AACAGGCACC GGCTCTGCTG CCTGCCGTCC AGCCGGACCC
1561 ACCTCTCCAC CAAGGAAGCT TCCAGTGTGC ACATGGGGAT TCCACCCTCC AATAAGGTGA
1621 TCAAGACCCG CTACCGCATT GTTAAGAAGA CCCCAAGCTC TTCCTTTGGT GCTCCATCCT
1681 TCCCCTCATC TCTACCCTCC TGGCGGGCCC GGCGCATCCC ATTATCCAGG TCCCTAGTGC
1741 TAAACCGCCT TCGTCCAGCA ATCACTGGGG GAGGGAAAGC CCCACCTGGT ACCCCTCGAT
1801 GGCGCAACAA AGGCTACCGC TGCATTGGAG GGGTTCTGTA CAAGGTGTCT GCCAACAAGC
1861 TCTCCAAAAC TTCTAGCAGG CCCAGTGATG GCAACAGGAC CCTCCTCCGC ACAGGACGCC
1921 TGGACCCTGC TACCACCTGC AGTCGTTCCT TGGCCAGCCG GGCCATCCAG CGGAGCCTGG
1981 CTATCATCCG GCAGGCGAAG CAGAAGAAAG AGAAGAAGAG AGAGTACTGC ATGTACTACA
2041 ACCGCTTTGG CAGGTGTAAC CGTGGCGAAT GCTGCCCCTA CATCCATGAC CCTGAGAAGG
2101 TGGCCGTGTG CACCAGATTT GTCCGAGGCA CATGCAAGAA GACAGATGGG TCCTGCCCTT
2161 TCTCTCACCA TGTGTCCAAG GAAAAGATGC CTGTGTGCTC CTACTTTCTG AAGGGGATCT
2221 GCAGCAACAG CAACTGCCCC TACAGCCATG TGTACGTGTC CCGCAAGGCT GAAGTCTGCA
2281 GTGACTTCCT CAAAGGCTAC TGCCCATTGG GTGCAAAGTG CAAGAAGAAG CACACGCTGC
2341 TGTGTCCTGA CTTTGCCCGC AGGGGTATTT GTCCCCGTGG CTCCCAGTGC CAGCTGCTCC
2401 ATCGTAACCA GAAGCGACAT GGCCGGCGGA CAGCTGCACC TCCTATCCCT GGGCCCAGTG
```

```
2461  ATGGAGCCCC CAGAAGCAAG GCCTCAGCTG GCCACGTACT CAGGAAGCCT ACTACTACTC
2521  AGCGCTCTGT CAGACAGATG TCCAGTGGTC TGGCTTCCGG AGCTGAGGCC CCAGCCTCCC
2581  CACCTCCCTC CCCAAGGGTA TTAGCCTCCA CCTCTACCCT GTCTTCAAAG GCCACCGCTG
2641  CCTCCTCTCC TTCCCCCTCT CCCTCTACTA GCTCGGGAGC CCCTTCCTTG GAGCAGGAAG
2701  AAGCTGTCTC TGGGACAGGC TCAGGAACAG GCTCCAGTGG CCTCTGCAAG CTGCCATCCT
2761  TCATCTCCCT GCACTCCTCC CCAAGCCCAG GAGGACAGAC TGAGACTGGG CCCCAGGCCC
2821  CCAGGAGCCC TCGCACCAAG GACTCAGGGA AGCCGCTACA CATCAAACCA CGCCTGTGAG
2881  GCCCCCTGAG GACCAGCCCG CACCTACCTC AGACCCTCAC CCCTGGAGAG GATGAAGGCT
2941  CTACCCACAA CTGCTCCTG
```

SEQ ID NO 4

```
  1 MetGluGluLysGlu GlnLeuArgArgGln IleArgLeuLeuGln
 16 GlyLeuIleAspAsp TyrLysThrLeuHis GlyAsnGlyProAla
 31 LeuGlyAsnSerSer AlaThrArgTrpGln ProProValPhePro
 46 GlyGlyArgThrPhe GlyAlaArgTyrSer ArgProSerArgArg
 61 GlyPheSerSerHis HisGlyProSerTrp ArgLysLysTyrSer
 76 LeuValAsnGlnPro ValGluSerSerAsp ProAlaSerAspPro
 91 AlaPheGlnThrSer LeuArgSerGluAsp SerGlnHisProGlu
106 ProGlnGlnTyrVal LeuGluArgGlnVal GlnLeuSerProAsp
121 GlnAsnMetValIle LysIleLysProPro SerLysSerGlyAla
136 IleAsnAlaSerGly ValGlnArgGlySer LeuGluGlyCysAsp
151 AspProSerTrpSer GlyGlnArgProGln GlySerGluValGlu
166 ValProGlyGlyGln LeuGlnProAlaArg ProGlyArgThrLys
181 ValGlyTyrSerVal AspAspProLeuLeu ValCysGlnLysGlu
196 ProGlyLysProArg ValValLysSerVal GlyArgValSerAsp
211 SerSerProGluHis ArgArgThrValSer GluAsnGluValAla
226 LeuArgValHisPhe ProSerValLeuPro HisHisThrAlaVal
241 AlaLeuGlyArgLys ValGlyProHisSer ThrSerTyrSerGlu
256 GlnPheIleGlyAsp GlnArgAlaAsnThr GlyHisSerAspGln
271 ProAlaSerLeuGly ProValValAlaSer ValArgProAlaThr
286 AlaArgGlnValArg GluAlaSerLeuLeu ValSerCysArgThr
301 SerLysPheArgLys AsnAsnTyrLysTrp ValAlaAlaSerGlu
316 LysSerProArgVal AlaArgArgAlaLeu SerProArgThrThr
331 LeuGluSerGlyAsn LysAlaThrLeuGly ThrValGlyLysThr
346 GluLysProGlnPro LysValAspProGlu ValArgProGluLys
361 LeuAlaThrProSer LysProGlyLeuSer ProSerLysTyrLys
376 TrpLysAlaSerSer ProSerAlaSerSer SerSerSerPheArg
391 TrpGlnSerGluAla GlySerLysAspHis ThrSerGlnLeuSer
406 ProValProSerArg ProThrSerGlyAsp ArgProAlaGlyGly
421 ProSerSerLeuLys ProLeuPheGlyGlu SerGlnLeuSerAla
436 TyrLysValLysSer ArgThrLysIleIle ArgArgArgGlyAsn
451 ThrSerIleProGly AspLysLysAsnSer ProThrThrAlaThr
466 ThrSerLysAsnHis LeuThrGlnArgArg ArgGlnAlaLeuArg
481 GlyLysAsnSerPro ValLeuArgLysThr ProHisLysGlyLeu
496 MetGlnValAsnArg HisArgLeuCysCys LeuProSerSerArg
511 ThrHisLeuSerThr LysGluAlaSerSer ValHisMetGlyIle
526 ProProSerAsnLys ValIleLysThrArg TyrArgIleValLys
541 LysThrProSerSer SerPheGlyAlaPro SerPheProSerSer
556 LeuProSerTrpArg AlaArgArgIlePro LeuSerArgSerLeu
571 ValLeuAsnArgLeu ArgProAlaIleThr GlyGlyGlyLysAla
586 ProProGlyThrPro ArgTrpArgAsnLys GlyTyrArgCysIle
601 GlyGlyValLeuTyr LysValSerAlaAsn LysLeuSerLysThr
```

616 SerSerArgProSer AspGlyAsnArgThr LeuLeuArgThrGly
631 ArgLeuAspProAla ThrThrCysSerArg SerLeuAlaSerArg
646 AlaIleGlnArgSer LeuAlaIleIleArg GlnAlaLysGlnLys
661 LysGluLysLysArg GluTyrCysMetTyr TyrAsnArgPheGly
676 ArgCysAsnArgGly GluCysCysProTyr IleHisAspProGlu
691 LysValAlaValCys ThrArgPheValArg GlyThrCysLysLys
706 ThrAspGlySerCys ProPheSerHisHis ValSerLysGluLys
721 MetProValCysSer TyrPheLeuLysGly IleCysSerAsnSer
736 AsnCysProTyrSer HisValTyrValSer ArgLysAlaGluVal
751 CysSerAspPheLeu LysGlyTyrCysPro LeuGlyAlaLysCys
766 LysLysLysHisThr LeuLeuCysProAsp PheAlaArgArgGly
781 IleCysProArgGly SerGlnCysGlnLeu LeuHisArgAsnGln
796 LysArgHisGlyArg ArgThrAlaAlaPro ProIleProGlyPro
811 SerAspGlyAlaPro ArgSerLysAlaSer AlaGlyHisValLeu
826 ArgLysProThrThr ThrGlnArgSerVal ArgGlnMetSerSer
841 GlyLeuAlaSerGly AlaGluAlaProAla SerProProProSer
856 ProArgValLeuAla SerThrSerThrLeu SerSerLysAlaThr
871 AlaAlaSerSerPro SerProSerProSer ThrSerSerProAla
886 ProSerLeuGluGln GluGluAlaValSer GlyThrGlySerGly
901 ThrGlySerSerGly LeuCysLysLeuPro SerPheIleSerLeu
916 HisSerSerProSer ProGlyGlyGlnThr GluThrGlyProGln
931 AlaProArgSerPro ArgThrLysAspSer GlyLysProLeuHis
946 IleLysProArgLeu

SEQ ID NO 5

1 GAGGCTTCGA AAGGTGCTGA AGCAGATGGG AAGGCTGCGC TGCCCCCAAG AGGGCTGTGG
61 GGCTGCCTTC TCCAGCCTCA TGGGTTATCA ATACCACCAG CGGCGCTGTG GGAAGCCACC
121 CTGTGAGGTA GACAGTCCCT CCTTCCCCTG TACCCACTGT GGCAAGACTT ACCGATCCAA
181 GGCTGGCCAC GACTATCATG TGCGTTCAGA GCACACAGCC CCGCCTCCTG AGGATCCCAC
241 AGACAAGATC CCTGAGGCTG AGGACCTGCT TGGGGTAGAA CGGACCCCAA GTGGTCGCAT
301 CCGACGTACG TGCCCAGGTT GCCGTGTTCC ATCTACAGGA GATTGCAGAG ATGAACTGGC
361 CCGTGACTGG ACCAAACAAC GCATGAAGGA TGACTTGTGC CTGAGAATGC ACGACTCAAC
421 TACACTCGGC CAGGTCTCCC CACACTTAAC CCTCAGCTGC TGGAAGCATG GAAGAATGAA
481 GTCAAGGAGA AGGGCCATGT GAACTGTCCC AATGAATTGC TGTGAAGCCA TCTACGCCAG
541 TGTGTCCGGC CTCAAGGCCC ATCTTGCCAG CTGCAGCAAG GGGGACCACC TGGGTGGGGA
601 AAGTACCGCT GCCTGCTGTG TCCCAAAGAA GTTCAGCTCT GAAAAGCGGC GTGAAGTTAC
661 CACATCCTTA AAGACCCAAC GGGAGAGAAT TGGTTCCGGA CCTCAGCTGA CCCGTCTTCC
721 AACACAAGAG CCAGGACTCC TTGATGCCTA GGAAAGAGAA AGAAATTTGT CAGGGAGAAA
781 GAAGCGGGGC CGCAAACCCA AGGAACGATC CTCCGAGGAG CCAGCATCTG CTCCCCCCTA
841 ACAGGGAATG ACTGGCCCCC AGGAGGCAGA GANAGGGGGT CCCGGAGCTC CACTGGGAAG
901 AAGGCTGGAG CTGGGAAGGC ACCTGAAAAG TGAGCCTAGT GGGCAGGGCC TACCCATCAT
961 GCCCTGCATT GTCCAGATTA GGGGAGCCAG TTCTAGACTG GTCCTCCACC TCCAACACAC
1021 ACCCCCATCT GTCCAGAGGG TTGGCAAACT ACTCTGCTCT CCCTGAAAGT GGTCCTTCCC
1081 CTGTTTAGGC TGCCTCAACA AGGCTAGATG GGGCTCCCCG GGAGTGCCAG GGCAGCAGCA

1141 AAAGTGCAAT AGGCTGGAGG ACCCAGCCGT TCCTACAAGG ACATTGCATG GCAGGAGCCT
1201 TGGCATCATG GGGCATGAAG TGTGCTTAAA CAGTTAAAAG GTCCCAGTTT CCACCTTCCT
1261 CTGGCCCAGT AGGATCCCCA ATCTGACTCT TTCAAGGCTC AGACATTCCT GGTGACCCAA
1321 TGTTGTGGAC TGATGAGGCA CCTGAGCAGT CTGGCTGCCA TAACTTGGGC CTCGCCTCCA
1381 CCCAACACTG GAACTCCAGT ACTCCCGGA

SEQ ID NO 6

1 GGATTTACTG CTCAGCCAGC TACTTACCAT CATAACAGCA CTACCACCTG GACTGGAAGT
61 AGGACTGCAC CATACACACC TAATTTGCCT CACCACCAAA ACGGCCATCT TCAGCACCAC
121 CCGCCTATGC CGCCCCATCC TGGACATTAC TGGCCAGTTC ACAATGAGCT TGCATTCCAG
181 CCTCCCATTT CCAATCATCC TGCTCCTGAG TACTGGTGCT CCATTGCTTA CTTTGAAATG
241 GACGTTCAGG TAGGAGAGAC GTTTAAGGTC CCTTCAAGTT GCCCTGTTGT GACTGTGGAT
301 GGCTATGTGG ATCCTTCGGG AGGAGATCGC TTTTGCTTGG GTCAACTCTC CAATGTCCAC
361 AGGACAGAAG CGATTGAGAG AGCGAGGTTG CACATAGGCA AAGGAGTGCA GTTGGAATGT
421 AAAGGTGAAG GTGACGTTTG GGTCAGGTGC CTTAGTGACC ACGCGGTCTT TGTACAGAGT
481 TACTACCTGG ACAGAGAAGC TGGCCGAGCA CCTGGCGACG CTGTTCATAA GATCTACCCA
541 AGCGCGTATA TAAAGGTCTT TGATCTGCGG CAGTGTCACC GGCAGATGCA GCAACAGGCG
601 GCCACTGCGC AAGCTGCAGC TGCTGCTCAG GCGGCGGCCG TGGCAGGGAA CATCCCTGGC
661 CCTGGGTCCG TGGGTGGAAT AGCCCCAGCC ATCAGTCTGT CTGCTGCTGC TGGCATCGGT
721 GTGGATGACC TCCGGCGATT GTGCATTCTC AGGATGAGCT TTGTGAAGGG CTGGGGCCCA
781 GACTACCCCA GGCAGAGCAT CAAGGAAACC CCGTGCTGGA TTGAGATTCA CCTTCACCGA
841 GCTCTGCAGC TCTTGGATGA AGTCCTGCAC ACCATGCCCA TTGCGGACCC ACAGCCTTTA
901 GACTGAGATC TCACACCACG GACGCCCTAA CCATTTCCAG GATGGTGGAC TAATGAAATA

SEQ ID NO 7

1 TTTTTTTTTT TCCACTTCGT ATAGTGACTC AGTTTTATTT ACGCTAGTAA CTAGGTAGAA
61 AGTATACATG TGTGTCTGTG GTACAGTCAA TGTGTCTTAA CTCCTCCACT TCAATCTCTA
121 CAAAGTCACC GCCAAGTGAT CAAGGATGGC AAACACAGGG CTTATAACCA AAAGGTATAA
181 AAAAGTCTGC AGTCTTGCCC TAAGATACAA AAACTGAATT TTAAACAATG TCAAAACATA
241 CATGATTTTA ACAAGTATAT GNAAAAGAAT CACACATCAA ATCAAGTACA AAAATATCCA
301 AACCACCTGT TACAACTGCA CTGTTTCCAT TATCCTGCAC AGTATTTAAC ATAAAAATTT
361 AGCAGTTTCC AAAAATATTC ATTAATTCAC TTGAAGTTAC TGCCCCNTGC AAAACAGTGA
421 AACACCAGGC AAACCAANCT GCCTTTAATT NTTTTNNACC AAATCNTCCT CCCNAN

SEQ ID NO 8

```
  1  GACAGAACCG GTTCGCACCG ACAGACGGAC AGAGGACCAG ACAGCCACTA AGGAGCGCTT
 61  ACTGCCCCCC TCCGGGCCCC TGCCCCGAAC TCCAGCCCCA GCGCCTGTTA CTGCCCCAGA
121  TACAGCAAGA TGCGCGGTCC TGGCAGCGAG ACACGGGCGA GCACTGTCCC CCGGTCCCCG
181  AGCCCTGGCC CCTAGCGCCC AGCGCTGCTG CCCTGCATCA GGGAGGGCCG CGGAGACCCC
241  AGCCTCAGTT GGCGCAGGAG CCCTGCGGGT GGGGCCTGCC CAGCCCAGCC AGGCGCGCCA
301  GCCCACCATG CTCCTCCTGT CGCCGCGCAG CGCGCTGGTC TCCGTCTATT GCCCGCAGAT
361  CTTTCTCCTT CTGTCCACGG CAGTTACTAC ATTGTCATCC GTGGTGGCCC TGGGAGCCAA
421  CATCATCTGC AACAAGATTC CTGGCCTGGC CCCACGGCAG CGTGCCATCT GCCAGAGCCG
481  ACCCGATGCC ATCATTGTGA TCGGGGAGGG GGCGCAGATG GGCATCGACG AGTGCCAGCA
541  CCAGTTCCGA TTCGGCCGCT GGAACTGCTC CGCCCTGGGC GAGAAGACCG TCTTCGGGCA
601  AGAACTCCGA GTAGGGAGTC GAGAGGCTGC CTTCACCTAT GCCATCACGG CGGCGGGCGT
661  GGCGCATGCT GTCACCGCTG CCTGCAGCCA GGGCAATCTG AGCAATTGTG GCTGTGACCG
721  GGAGAAGCAA GGCTACTACA ACCAGGCGGA AGGCTGGAAG TGGGGGGGCT GCTCAGCGGA
781  CGTCCGCTAC GGCATCGACT TTTCTCGTCG CTTTGTGGAT GCCCGTGAGA TCAAAAAGAA
841  CGCCGGATCC
```

SEQ ID NO 9

```
1    AGACACTGTT GTATTCAGAT TATTTCTTAG TGGCTGGCTT TTGATTCTAG ACAGAGATTC
61   TTAAAGTCCT TTTAAAAAAG TGGATCAGGA ATCCTGTTAT GGGCCTTGAT TGTTCCAGAC
121  ATTAGAAGTA AATATATTTG ATGAAGGAAA TCTTGAAAAA ATACTGACTA GATAAAAATT
181  GTAAGCCAAG CTTTCTGACT GAAAAATGCT ACCTAGCCAC AGATCATTGC TGTTATTTGG
241  TTCATTGCAT GAGTGTGTAT GTGTGTGTAT ATATGTATAC ACATATATAT GTGTGTGTGT
301  GTGTATGTGT ACACACACAT ATATGTGGGT TTTGGGGGGT ATGGATAAGA TGGTGCTATG
361  AAAATAATTT GTCTCTTGTT TTAATTAATG AAGCTTCTGT CATGCCAAGT AATCTTTAAG
421  GGAGAATCAG AACTTTTCAT TAAAANTCAT AAGGGAAACA GAATTTGTAC GGGTG
```

SEQ ID NO 10

```
  1  AGCGGAGTTT CAGTCTGCGG ACACGCGTGG AGCCCTTGCC CGGGCCTCCG TGGGTCTGAG
 61  GCGCTGCGAG CCCTGGGTAA CCACGGCCTC GAGCTGCTGT CCTCACCAAG ATCCTCCAAT
121  TCTGAACCAA GAACAAAAAA ATGTTTCAGC TTCGTGCATT TCAAAGAAGG CATTAACTAG
181  AGCCCAGTTT GGCGGACAAG TTCTTCATTC AAAAGAGAGT CCTGTTAGGA TCACTGTGTC
241  CAAAAAGAAC ACATTTGTTT TGGGAGGCAT TGATTGTACT TATGAAAAGT TTGAAAATAC
301  TGATGTTAAC ACCATTAGTT CTCTTTGTGT TCCTATTAAG AATCATAGCC AATCTATTAC
361  TTCTGATAAT GATGTGACAA CAGAAAGGAC TGCAAAAGAG GATATTACAG AACCAAATGA
421  AGAGATGATG TCCAGAAGAA CTATTCTTCA AGATCCCATA AAGAATACAT CTAAAATTAA
481  ACGTTCAAGT CCAAGACCTA ATTTAACACT ATCTGGCCGG TCTCAAAGAA AATGTACAAA
541  GCTTGAAACT GTTGTAAAAG AAGTAAAAAA ATATCAGGCA GTCCACCTAC AGGAATGGAT
601  GATTAAAGTC ATCAATAATA ATACTGCTAT ATGTGTAGAA GGAAAGCTGG TAGATATGAC
661  TGATGTTTAT TGGCATAGCA ATGTAATTAT AGAGCGGATT AAACACAATG AACTTAGGAC
721  CTTATCAGGC AACATTTATA TCTTAAAAGG ATTGATAGAC TCGGTCTCCA TGAAAGAAGC
781  AGGATATCCC TGTTATCTCA CAAGAAAATT TATGTTTGGA TTTCCCCACA ACTGGAAGGA
841  ACACATTGAT AAATTTCTAG AACAATTAAG GGCTGAAAAA AAGAACAAGA CCAGACAGGA
901  AACAGCAAGA GTCCAAGAAA AACAAAAATC AAAAAAAAAA GATGCAGAAG ATAAAGAAAC
```

961 TTATGTCCTC CAAAAGGCCA GCATCACGTA TGACCTTAAT GATAATAGCT TAGAGAGAAC
1021 TGAAGTACCC ACTGATCCCT TGAACTCACT GGAACAGCCT ACCTCCGGCA AAGAAAGAAG
1081 ACACCCGCTT CTCAGTCAGA AGAGAGCTTA TGTTTTAATA ACACCACTTA GAAACAAAAA
1141 GTTGATAGAG CAAAGATGTA TAGACTACAG TCTGTGTATT GAAGGAATAT CGGACTTTTT
1201 CAAAGCAAAG CATCAAGAAG AAAGTGACTC AGATATACAT GGAACTCCAA GTTCTACCAG
1261 TAAGTCTCAA GAGACCTTTG AACATAGAGT GGGATTTGAA GGCAATACCA AGGAGGACTG
1321 CAATGAATGT GACATAATCA CTGCCAGACA TATTCAGATA CCTTGCCCGA AAAGTAAACA
1381 AATGCTCACC AATGATTTTA TGAAAAAGAA CAAGTTGCCC TCAAAACTGC AGAAAACTGA
1441 AAATCAAATA GGTGTATCAC AGTATTGCCG GTCCTCATCA CATTTGTCAA GTGAAGAGAA
1501 TGAAGTAGAA ATTAAAAGTA GAACCAGAGG ATCCCAA

SEQ ID NO 11

1 GAGTAAACTC TCCTTCCGAG CGCGGGCGCT GGACGCCGCC AAACCGCTGC CCATCTACCG
61 CGGCAAGGAC ATGCCTGATC TCAACGACTG CGTCTCCATC AACCGGGCCG TGCCCCAGAT
121 GCCCACCGGG ATGGAGAAGG AGGAGGAATC GGAACATCAC CTACAGCGAG CTATTTCAGC
181 GCAGCAAGTA TTTAGAGAAA AAAAAGAGAG CATGGTCATT CCAGTTCCTG AGGCAGAGAG
241 CAACGTCAAC TATTACAATC NGCTTGTACA AAGGGGAGTT CAAACAGCCC AAGCAGTTCA
301 TNCATATTCA GCCTTTTAAC CTAGACAACG AGCAACCAGA TTATGATATG GATTCAGAAG
361 ATGAGACATT ATTAAATAGA CTTAACAGAA AAATGGAAAT TAAACCTTTG CAATTTGAAA
421 TTATGATTGA CAGACTTGAA AAAGCCANTT CTACCAGCTT GTACACTTCA AGAAGCA

SEQ ID NO 12

1 TCTGGTTCTA CTTTTAATTT CTACTTCATT CTCTTCACTT GACAAATGTG ATGAGGACCG
61 GCAATACTGT GATACACCTA TTTGATTTTC AGTTTTCTGC AGTTTTGAGG GCAACTTGTT
121 CTTTTTCATA AAATCATTGG TGAGCATTTG TTTACTTTTC GGGCAAGGTA TCTGAATATG
181 TCTGGCAGTG ATTATGTCAC ATTCATTGCA GTCCTCCTTG GTATTGCCTT CAAATCCCAC
241 TCTATGTTCA AAGGTCTCTT GAGACTTACT GGTAGAACTT GGAGTTCCAT GTATATCTGA
301 GTCACTTTCT TCTTGATGCT TTGCTTTGAA AAATCCGATA TTCCTTCAAT AGAGAGACTG
361 TAGTCTATAC ATCTTTGCTC TATCAACTTT TTGTTTCTAA GTGGTGTTAT TAAAACATAA
421 GCTCTCTTCT GACTGAGAAG CGGGTGTCTT CTTTCTTTGC CGGAGGTAGC TGTTCCAGTG
481 ATTCAAGGGA TCAATGGGTA CTCANTCTCT CTAANCTATA TCATAAGGTC TACTTAATGC
541 TGGCTTTTGG AAGANTAATT CTTTATCTCT GN

SEQ ID NO 13

1 CTGCTGTGAG GAATGCTGGG ATTGTTGTTT CTGATGAAGC TGCGCAAGTT GCTGCCTTTG
61 CATTTGAACT AGCTGCTGTT GATGTGTCTG AAACTGCTCT TCTGTGATGC CCCCTGTTAC
121 TGATATGCCG TTCTTGCTGG TGTTCAATAA AGCTACGGAT GCTGCAGAAA CTCTTTTACT

181 GCTCACAGTC TGCCCTGGTT TTCTTGAGGT ACATTCTTCA CTATCAATGT CCTGTACATT
241 TAGTAGCCTT GGCTGGAAAC ACTGTAGTCG ACATGATCTG ATATTGCTTA ATATTTCAGA
301 AAGAGACAGT CTATNTTCAC AAGGTTTACT GGGAAGCATT GGTCCGAGAG AAATTAGAAG
361 AAAATCTATA GTTTGGGAAG ACTTGAAAAC CCGTTCAGCA TCTCANGGTC TATCTGTTTC
421 AGGACGGGGT CATGTTCTGT GGATATCCGT CCATTATGAA CCTGCCACTC TGCCATTCCC
481 CTCCTTGCAA TCCTATACAT CTTCTTGGAC TGTAATTTCG TAAGANATGC TTATACTCAA
541 CTTATCCAAT CTGCCACTCT GAATTTCNAC ATATGGTAN

SEQ ID NO 14

1 GGAAAGACAA AGATGCAGGA TATAGTACTT GGAACAGGCT TTTTAAGTAT TCATCCTAAA
61 AATGAGGCTG AGCACATAGA AAATGGGGCT AAGTGTCCGA ATTTGGAGTC CATAAATAAG
121 GTAAATGGTC TTTGTGAGGA CACTGCACCG TCTCCTGGTA GGGTTGAACC ACAGAAGGCC
181 AGTTCTTCTG CTGACGTGGG CATTTCTAAA AGCACGGAAG ATCTATCTCC TCAGAGAAGT
241 GGTCCAACTG GAGCTGTTGT GAAATCTCAT AGTATAACTA ACATGGAGAC TGGAGGCTTA
301 AAAATCTATG ACATTCTTGG TGATGATGGC CCTCAGCCGC CAAGTTGCAG CAGTTAAAAT
361 CGCATCTGCT GTGGATGGGG AAGAACATAT CAGAAGCAAN TCT

SEQ ID NO 15

1 TTTTTTTTTT TTTTTTTTTT GACAGTTTTG AAATTATATT TATTAATGCT TTATTATACG
61 TATTGTATTC TATTTGAGCC AAGGGAAAGG AGAACCCCAC TCAAGTGAGA TAACAAACTT
121 GCTGTCTTTT ACAAAATTTA ATCAGAACTG ACAATGTTAT GGTTAGTTCT TAATTCCTGA
181 GAATTTGAAC ATCATTAAGT TTTCTGTGAA TTTACAACAA AACACTCATG TTAATATTTA
241 AATTACAATA TTTCTGAAAA AATATTGTTA GCAAAAGAAA ACCACATCCA ACGTATACAG
301 TAACCCAGGT GTGAACATAC TGAAGCCCTG TTGCTCAGCA GTTTAATACC ATTTAAATAT
361 TTCTCTCATC AGAGATTTAT TNCAAATACA TGAACTTATT ATAATTTACC AGAATACAGT
421 GACATNATTT TTNTTTTTTT TTAAANAATT ATTATCTATT ATATGTAAGT ACCCGGTANC
481 TGTCTTCAAC ACCCAGAANA AGGGGTCCAA TCTTTTACAG AAGGTGTGAC CNCATGTGGN
541 GNCGGGAATT NANNN

SEQ ID NO 16

1 CTACGAAATT GTACCTGAGT GACATAAACC GGTAAAGGTG TGTTACTTCG CTTTTTCATG
61 TTTTTTTTTT CTTTTTGTTC TTTGGTCTGA TAAGAAAATG GACAGTTGTG GAAAGTCAGG
121 TAATACAGAT CAGTTTCCAG TTCAGAACCC TAAATCACAC CTACGTGAGT GAGGCTGCTG
181 CACTGCTTTC CTTGGGTTCT TCGGCCGGCC AGACAGCCTT TCTGCTTTGT AAGTGACTTC
241 ATTATAGCCA TCAGCTAATC ACTCCCTCAG CATACACTGG CATCTCCAGA TTACCTGACG
301 GCAGACATAC TTGCTCTGGC TTCAATTAAC ATGCTGTCAA GCATCCCTCT CGACATTCAC
361 ATGGCAACAC AAAACCATGA ATTTCTCTTC ATACAACCAG GAATACACAC TCATAAAGGG

421 AAAGCGTTAN ACCTGATTTT TATTAAATAT TATTTCCTTC CCTTTCCATG CCAAGTTCAC
481 GTTAACATCT TTAGAATACT AAAACGGAAA CCCNCCACTT ANGAAACAAC TGGGAATTGG
541 ACATCCACAG GTACATCACA NA

SEQ ID NO 17

1 AGCGGNAGTT TCAGTCTGCG NGACACGCGT GGNAGCCCTT GCCCGGGCCT CCGTGGGTCT
61 GAGGCGCTGC GAGCCCTGGG TAACCACGGC CTCGAGCTGC TGTCCTCACC AAGATCCTCC
121 AATTCTGAAC CAAGAACAAA AAAATGTTTC AGCTTCGTGC ATTTCAAAGA AGGCATTAAC
181 TAGAGCCCAG TTTGGCGGAC AAGTTCTTCA TTCAAAAGAG AGTCCTGTTA GGATCACTGT
241 GTCCAAAAAG AACACATTTG TTTTGGGAGG CATTGATTGT ACTTATTGAA AAGTTTGAA
301 AATACTGATG TTTAACACCA TTAAGTTCTC TTTGTGTTNC CTAATTA

SEQ ID NO 18

1 CCTCAATGTG TCGTAGTACT TGTTCCCGCC AGTCATGAGG AACCTTGCTT TTTCCTGGAG
61 GATCTAACAG AGAATGTTCA GACCCGACCC TTGTATTTGG TCTTTTTGAA GGACTAGTCC
121 GTGAGTAATT GAAATCACTA ACTGACATAG TTCTCNCNGN TATTTCATTA ATAGAGGGAC
181 GGGCACTCTG AGGCCTGGAT GTATTTGGGC CATCGATGCT GTACGCTCGT GCAGAAAGAG
241 GTCTCTGTGA TCCTGACATG ACTGGAGTTC TTCCCATTGA ATGTAACTCT CTGTACGATA
301 AGTAATCTCC TTCAGTACGC CTTGTGGGGT CACCGAGATT TACAGAAGCC GTTGAAGACA
361 CGCTACTCTG TCTCTGAATA GTAATCCGAA TGACTGCTGG CACTAGTCGG TCATTCNGGG
421 AGATACCCAC ATTTCTCCAT GCCTGGCTGG GGCAATCTCT GTTGTAANTG GTATCCAATA
481 TTGGTCTACA TTGTTATGGT TAAAAAAATC TGTTTGGAGA ATGCTTTGCA TACTGTNAAT
541 TTCTGCCTCN CAAATNTTGG AAGGNCCGA

SEQ ID NO 19

1 GAGACATTCT GAAGGGCAGG AATGAGGCGC TCTCCCCAGG GNAGATGGTG GTGAGGCTGC
61 TGAGGGGGAA GGTGATATCT TTCCATCTTC TCATTACCTG CCAATCACCA AAGAAGGCCC
121 TCGAGACATT CTGGATGGCA GAAGTGGCAT TTCTGTGGCT AACTTCGACC CGGGCACCTT
181 TAGCCTGATG CGATGTGACT TCTGTGGGGC TGGTTTTGAT ACTCGGGCTG GCCTCTCCAG
241 TCATGCCCGG GCCCACCTTC GTGACTTTGG CATCAGGAAC TTGGGGAACT CCACCATCTC
301 ACCATCAACA TCCTTGCAAA NAACTTGCTG GGCCACCT

SEQ ID NO 20

```
1   GGAGGGTGTA GCAAGGCCTG AGAACATCTT CCGGGCCGTG GGAGGAGGAG AAGCAGTTGG
61  TGAGTGGCCC AGAGGACTGC CTGGTGGTGG TGGCAACTTC TTGGTCAAAG GTGAGATGTG
121 AAGATCAGAG GGACTTCGGG CTTCTAGTGA GCTGCCAGGA CCTCCAGTGC TCAGCACCTT
181 GGCCAGGGCT TTTGGGCTAG GACCTGGTGG GTGGAGGTGT CCCCCTGGCC TGGATTGGGT
241 CCGTCTCTTC AGGATCTCCC GAAGTGTGTC GATGGGTGAG CCGTTCACAT ACCACTCAGT
301 TACACCCATC TGGCGCANGT GGGAACGTGC ATGGCTANAC AAGCCCTTTC TGTTCTCAAA
361 GAATCACCAC ANAACTCACA GCGGATATCT CTTGTTGGCT CTGGGCCTGA ANCATCTCCG
421 TANATTGGCC CANGGTCCTC ACCCCANTTA NGCGGGAAAG GCATGGTNAA AAGTAACCTT
481 NGC
```

Claims

1. SMAD interacting protein(s) obtainable by a two-hybrid screening assay whereby Smad C-domain fused to a DNA-binding domain as bait and a vertebrate cDNA library as prey are used.

2. SMAD interacting protein (SIP) characterized in that:

   a) it fails to interact with full size XSmad1 in yeast
   b) it is a member of the family of zinc finger/homeodomain proteins including δ-crystallin enhancer binding protein and/or Drosophila zfh-1
   c) $SIP1_{czf}$ binds to E2 box sites
   d) $SIP1_{czf}$ binds to the Brachyury protein binding site
   e) it interferes with Brachyury-mediated transcription activation in cells
   f) it interacts with C-domain of Smad 1, 2 and/or 5

3. Isolated nucleic acid sequence comprising the nucleotide sequence as provided in SEQ ID NO 1 coding for a SMAD interacting protein or a functional fragment thereof.

4. A recombinant expression vector comprising the isolated nucleic acid sequence according to claim 3 operably linked to a suitable control sequence.

5. Cells transfected or transduced with a recombinant expression vector according to claim 4.

6. A nucleic acid sequence hybridizing to the nucleotide sequence as provided in SEQ ID NO 1 or part thereof and encoding a Smad interacting protein or a functional fragment thereof.

7. A polypeptide comprising the amino acid sequence according to SEQ.ID.NO 2 or a functional fragment thereof.

8. A pharmaceutical composition comprising a nucleic acid sequence according to claim 3 or claim 6.

9. A pharmaceutical composition comprising a polypeptide according to claim 7.

10. Method for diagnosing a disease by using a nucleic acid sequence according to claim 3 or claim 6.

11. Method for diagnosing a disease by using a polypeptide according to claim 7.

12. Method of screening for compounds which affect the interaction between SMAD and SMAD interacting protein.

13. Diagnostic kit comprising a nucleic acid sequence according to claim 3 or claim 6 and/or a polypeptide according to claim 7 for performing a method according to claim 10 or claim 11.

14. Transgenic animal harbouring the nucleic acid sequence of claim 3 or claim 6 in its genome.

15. Use of transgenic animal according to claim 14 for testing medicaments and therapy models.

16. Isolated nucleic acid sequence comprising the nucleotide sequence as provided in SEQ ID NO 3 coding for a SMAD interacting protein or a functional fragment thereof.

17. A polypeptide comprising the amino acid sequence according to SEQ.ID.NO 4 or a functional fragment thereof.

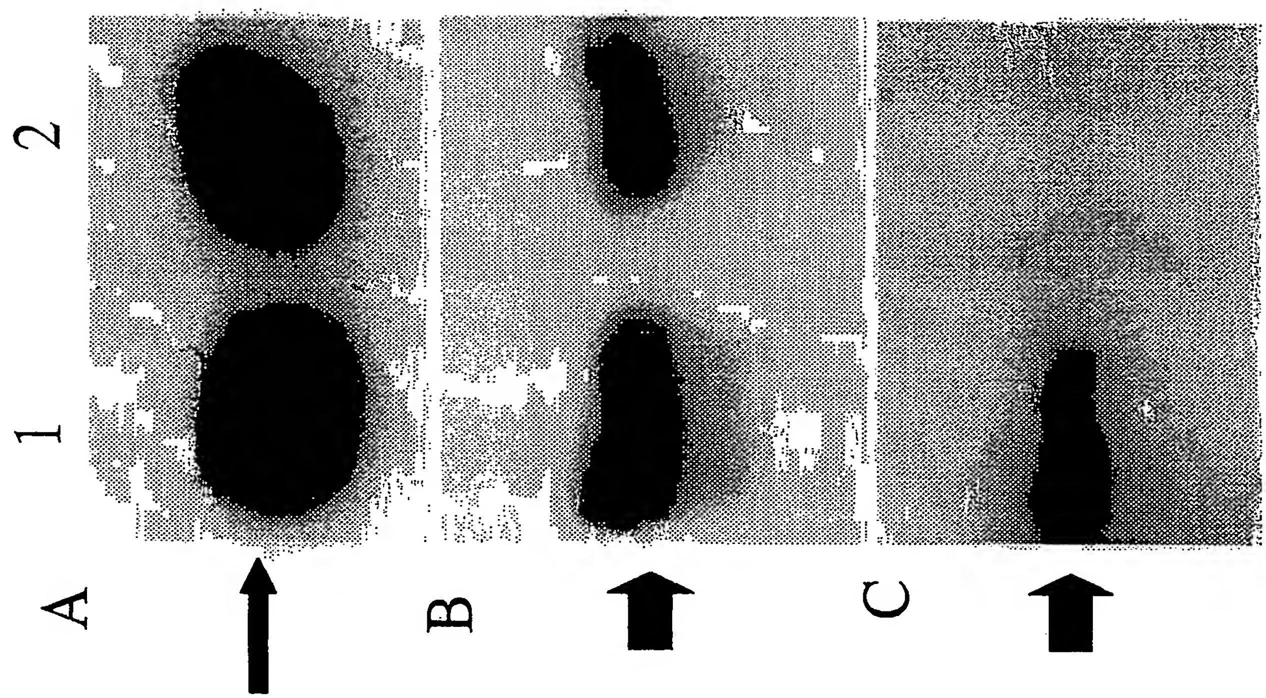
18. Isolated nucleic acid sequence comprising the nucleotide sequence as provided in SEQ ID NO 8 coding for a SMAD interacting protein or a functional fragment thereof.

19. Isolated nucleic acid sequence comprising the nucleotide sequence as provided in SEQ ID NO 10 coding for a SMAD interacting protein or a functional fragment thereof.

20. A polypeptide comprising the amino acid sequence depicted as the one letter code QHLGVGMEAPLLGFPTMNSNLSEVQKVLQIVDNTVSRQKMDCKTEDISKLK necessary for binding with Smad.

21. SMAD interacting protein characterized in that:
   a) it interacts with full size XSmad1 in yeast
   b) it is a member of a family of proteins which contain a cluster of 5 CCCH-type zinc fingers including Drosophila "Clipper" and Zebrafish "No arches"
   c) it binds single or double stranded DNA
   d) it has an RNase activity
   e) it interacts with C-domain of Smad1, 2 and/or 5.

22. A method for post-transcriptional regulation of gene expression by members of the TGF-β superfamily by manipulation or modulation of the interaction between Smad function and/or activity and mRNA stability.


A
B
C
1
2

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N 15/12, C07K 14/46, A61K 31/70, 38/17, C12Q 1/68, G01N 33/566, A01K 67/027

A3

(11) International Publication Number: WO 98/55512

(43) International Publication Date: 10 December 1998 (10.12.98)

(21) International Application Number: PCT/EP98/03193

(22) International Filing Date: 28 May 1998 (28.05.98)

(30) Priority Data:
97201645.5 2 June 1997 (02.06.97) EP
(34) Countries for which the regional or international application was filed: NL et al.

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

(71) Applicant *(for all designated States except US)*: VLAAMS INTERUNIVERSITAIR INSTITUUT VOOR BIOTECHNOLOGIE [BE/BE]; Rijvisschestraat 118, B-9052 Zwijnaarde (BE).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: VERSCHUEREN, Kristin [BE/BE]; Twee Leeuwenstraat 22, B-3078 Everberg (BE). REMACLE, Jacques [BE/BE]; Avenue des Lilas 7, B-4280 Hannut (BE). HUYLEBROECK, Danny [BE/BE]; Lijsterlaan 15, B-1770 Liede/kerke (BE).

(74) Common Representative: VLAAMS INTERUNIVERSITAIR INSTITUUT VOOR BIOTECHNOLOGIE; Rijvisschestraat 118, B-9052 Zwijnaarde (BE).

Published
With international search report.

(88) Date of publication of the international search report: 11 March 1999 (11.03.99)

(54) Title: SMAD-INTERACTING POLYPEPTIDES AND THEIR USE

(57) Abstract

The current invention concerns SMAD-interacting protein(s) obtainable by a two-hybrid screening assay whereby Smad1 C-domain fused to GAL4 DNA-binding domain as bait and a cDNA library from mouse embryo as prey are used. Some characteristics of a specific SMAD interacting protein so-called SIP1 are the follwing: a) it fails to interact with full size XSmad1 in yeast; b) it is a member of the family of zinc finger/homeodomain proteins including δ-crystallin enhancer binding protein and/or Drosophila zfh-1; c) $SIP1_{czf}$ binds to E2 box sites; d) $SIP1_{czf}$ binds to the Brachyury protein binding site; e) it interferes with Brachyury-mediated transcription activation in cells and f) it interacts with C-domain of Smad 1,2 and 5. The minimal length of the amino acid sequence necessary for binding with Smad appears to be a 51 aa domain encompassing aa 166-216 of SEQ ID NO 2 having the amino acid sequence as depicted in the one letter code: QHLGVGMEAPLLGFPTMNSNLSEVQKVLQIVDNTVSRQKMDCKTEDISKLK.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP 98/03193

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12N15/12 C07K14/46 A61K31/70 A61K38/17 C12Q1/68
G01N33/566 A01K67/027

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12N C07K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHEN X ET AL: "A transcriptional partner for MAD proteins in TGF-beta signalling" NATURE, vol. 383, no. 6602, 24 October 1996, pages 691-696, XP002047848 cited in the application see paragraph bridging pages 691 and 692	1,12
X	DATABASE EMBL EST 16: "Mus musculus cDNA clone 584313 5' DNA-binding protein" ACCESSION NO AA125512,26 November 1996, XP002084026 compare nucleotides 1-461 of AA125512 with nucleotides 1567-2027 in SEQ ID NO:1	3,6,7

-/--

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 November 1998	27/11/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Cupido, M

1

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	DE CAESTECKER M P ET AL.: "Characterization of functional domains within Smad4/DPC4" JOURNAL OF BIOLOGICAL CHEMISTRY, vol. 272, no. 21, 23 May 1997, pages 13690-13696, XP002084021 MD US see page 13690 ---	1
X	LAGNA G ET AL: "Partnership between DCP4 and SMAD proteins in TGF-beta signalling pathways" NATURE, vol. 383, no. 6603, 31 October 1996, pages 832-836, XP002047850 cited in the application see the whole document ---	1
X	MEERSSEMAN G ET AL: "The C-terminal domain of MAD-like signal transducers is sufficient for biological activity in the Xenopus embryo and transcriptional activation" MECHANISMS OF DEVELOPMENT, vol. 61, no. 1/02, January 1997, pages 127-140, XP002047849 cited in the application see the whole document ---	1
A	DATABASE EMBL HUM1: "Human mRNA for KIAA0150 gene, partial cds." ACCESSION NUMBER D63484,3 August 1996, XP002084022 compare nucleotides 1-2908 of D63484 with nucleotides 38-2952 in SEQ ID NO:3 ---	16,17,21
X	DATABASE EMBL EMROD: "Mouse Wnt-7b mRNA, completet cds." ACCESSION NUMBER M89802,3 April 1992, XP002084023 cited in the application compare nucleotides 74-529 in M89802 with nucleotides 391-848 in SEQ ID NO:8 ---	18
X	DATABASE EMBL EST16: "Stratagene mouse melanoma. Mus musculus cDNA clone 651678 5'" ACCESSION NUMBER AA212269,3 February 1997, XP002084024 cited in the application Comapar nucleotides 1-432 of AA212269 with nucleotides 930-1362 in SEQ ID NO:10 ---	19

-/--

1

International Application No
PCT/EP 98/03193

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
P,X	DATABASE EMBL EMHUM1: "Homo sapiens mRNA for KIAA0569 protein, complete cds." ACCESSION NUMBER AB011141,10 April 1998, XP002084025 compare nucleotides 1250-4249 in AB011141 with nucleotides 8-3007 in SEQ ID NO:1	1-7

1

International application No.

PCT/EP 98/03193

INTERNATIONAL SEARCH REPORT

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☒ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. Claims: 1-15, 20, 22

   Smad interacting proteins such as SIP1, encoding nucleotide sequences, corresponding pharmaceutical compositions diagnostic methods and transgenic animals, method for post-transcriptional regulation of gene expression by modulating Smad interaction.

2. Claims: 16, 17, 21

   Smad interacting proteins having the characteristics of SIP2, corresponding sequences.

3. Claim: 18

   Nucleic acid sequence encoding SIP7 or a functional fragment thereof.

4. Claim:19

   Nucleic acid sequence encoding SIP5 or a functional fragment thereof.